DOCUMENT RESUME

ED 173 433 TM 009 570

TITLE	Proceedings of the Invitational Conference on Testing Problems (16th, New York, New York, November 1, 1952)
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	1 Nov 52
NOTE	135p.
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Culture Free Tests; Elections; *Intelligence Tests; *Motivation; *Norm Referenced Tests; Prediction; Public Opinion; Student Testing; *Surveys; *Test Bias; Test Construction; *Testing Problems; Test Interpretation
IDENTIFIERS	Semantic Test of Intelligence

ABSTRACT

Four topics were emphasized during this conference on testing problems: (1) the selection of appropriate score scales for tests; (2) the experimental approach to the measurement of human motivation; (3) trends in public opinion polling since 1948 and their probable effects on predictions of the 1952 election; and (4) techniques for developing unbiased intelligence tests. Eric F. Gardner and Ledyard R. Ticker both discussed the importance of reference groups in scaling procedure; discussions by John C. Flanagan and E. F. Lindquist followed. David C. McClelland presented the address on the measurement of human motivation; a panel discussion followed. The luncheon symposium focused on the trends in public opinion polling since 1948, and the prediction of the 1952 elections. Sampling, interviewing, and data analysis were discussed by Frederick F. Stephan, Herbert Human, and Samuel Stouffer, respectively. The session on unbiased intelligence tests included papers by Irving Lorge, Phillip J. Rulon, and Ernest A. Haggard; Quinn McNemar and Ernest A. Haggard commented on the papers. (An example of Phillip Rulon's Semantic Test of Intelligence--STI--is included.) (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED173433

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

1952

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Educational
Testing Service
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM009 570

EDUCATIONAL TESTING SERVICE

EDUCATIONAL TESTING SERVICE

BOARD OF TRUSTEES

Katharine E. McBride, *Chairman*

Arthur S. Adams
Raymond B. Allen
Joseph W. Barker
Frank H. Bowles
Oliver C. Carmichael
Charles W. Cole
James B. Conant
Henry H. Hill
Herold C. Hunt
Lewis W. Jones
Thomas R. McConnell
Lester W. Nelson
Edward S. Noyes
George D. Stoddard

OFFICERS

Henry Chauncey, *President*
Richard H. Sullivan, *Vice President and Treasurer*
William W. Turnbull, *Vice President*
Jack K. Rimalover, *Secretary*
Catherine G. Sharp, *Assistant Secretary*
Robert F. Kolkebeck, *Assistant Treasurer*

COPYRIGHT, 1953, EDUCATIONAL TESTING SERVICE
20 NASSAU STREET, PRINCETON, N. J.
PRINTED IN THE UNITED STATES OF AMERICA

INVITATIONAL CONFERENCE ON TESTING PROBLEMS

NOVEMBER 1, 1952

GEORGE K. BENNETT, *Chairman*

¶ Selecting Appropriate Score Scales for Tests

¶ The Measurement of Human Motivation: An Experimental Approach

¶ Trends in Public Opinion Polling Since 1948 and Their Probable Effect on 1952 Election Predictions

¶ Techniques for the Development of Unbiased Tests

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY LOS ANGELES, CALIFORNIA

4

FOREWORD

The 1952 Invitational Conference on Testing Problems marked the sixteenth year of the Conference and the fifth year it has been sponsored by Educational Testing Service. With the transfer of the testing activities of the American Council on Education to the newly-formed Educational Testing Service in early 1948, it had seemed appropriate to transfer also the sponsorship of the thriving annual Conference. Under the Council's capable guidance the Invitational Conference had grown from a small group to almost two hundred interested participants.

This year the number of persons attending reached a new high, almost 100 more than the previous year and more than double that of five years ago. This would seem to be attributable not only to the growth of interest in measurement problems generally, but also to the particular appeal of the program arranged by Chairman George K. Bennett.

In planning his program Dr. Bennett reached coast to coast for the most able men to discuss the topics scheduled. For the luncheon symposium he arranged a program closely related to the national election which followed the Invitational Conference by three days. His efforts resulted in a meeting of great professional value and intellectual stimulation, one well befitting the history of successful Invitational Conferences.

To George Bennett and those participants who made this year's Conference so significantly successful I want to express my deep appreciation for a job well done.

Henry Chauncey
President

PREFACE

THE papers and discussions of the 1952 Invitational Conference on Testing Problems sponsored by Educational Testing Service are permanently recorded on the pages that follow. The Conference, held November 1, 1952, at the Roosevelt Hotel in New York City, attracted more than 400 individuals. There were four sections in the program, a morning panel discussing "Selecting Appropriate Score Scales for Tests," an address by Dr. David C. McClelland on "The Measurement of Human Motivation: An Experimental Approach," a luncheon symposium on "Trends in Public Opinion Polling Since 1948 and Their Probable Effect on 1952 Election Predictions," and an afternoon panel on "Techniques for the Development of Unbiased Tests."

As in past years the topics selected have been those regarded as timely in interest and important in psychometric implications. The quality of the audience at these meetings is such as to stimulate speakers to make carefully prepared and logical presentations of their points of view.

It is felt that the papers presented on this occasion have maintained the high level established by previous participants. It does not seem appropriate for the Chairman of this session to comment further upon the topics considered, but it is seemly for him to express his gratitude to Educational Testing Service for the privilege of presiding as well as his confidence that the Invitational Conference will exert a beneficial influence upon measurement in education and psychology for many years to come.

GEORGE K. BENNETT, CHAIRMAN
1952 Conference

CONTENTS

FOREWORD by Dr. Chauncey v

PREFACE by Dr. Bennett vii

PANEL I: *"Selecting Appropriate Score Scales for Tests"*

THE IMPORTANCE OF REFERENCE GROUPS IN SCALING PROCEDURE
Eric F. Gardner, *Syracuse University* 13

SCALES MINIMIZING THE IMPORTANCE OF REFERENCE GROUPS
Ledyard R Tucker, *Educational Testing Service* 22

DISCUSSION
John C. Flanagan, *American Institute for Research* 29

DISCUSSION
E. F. Lindquist, *State University of Iowa* 34

ADDRESS: *"The Measurement of Human Motivation: An Experimental Approach"*

David C. McClelland, *Wesleyan University* 41

DISCUSSION 52

LUNCHEON SYMPOSIUM: *"Trends in Public Opinion Polling since 1948 and Their Probable Effect on 1952 Election Predictions"*

WHAT ABOUT THE SAMPLING? A BIT OF PSEUDOHISTORY
Frederick F. Stephan, *Princeton University* 58

INTERVIEWING
Herbert Hyman, *Columbia University* 64

ANALYSIS
Samuel Stouffer, *Harvard University* 70

PANEL II: "*Techniques for the Development of Unbiased Tests*"

DIFFERENCE OR BIAS IN TESTS OF INTELLIGENCE
Irving Lorge, *Teachers College, Columbia University* 76

A SEMANTIC TEST OF INTELLIGENCE
Phillip J. Rulon, *Harvard University* 84

TECHNIQUES FOR THE DEVELOPMENT OF UNBIASED TESTS
Ernest A. Haggard, *University of Chicago* 93

DISCUSSION OF PAPERS
Quinn McNemar, *Stanford University* 121

REPLY TO DR. MCNEMAR'S REMARKS
Ernest A. Haggard, *University of Chicago* 125

DISCUSSION 129

APPENDIX 132

8

PANEL I

Selecting Appropriate Score Scales for Tests

Selecting Appropriate Score Scales for Tests

ERIC F. GARDNER

THE IMPORTANCE OF REFERENCE GROUPS IN SCALING PROCEDURE

IT IS commonly accepted that a single isolated test score is of little or no value. For a score to have meaning and be of social or scientific utility, some sort of frame of reference is needed. A number of different frames of reference have been proposed and been found to have value. In view of the fact that this session is devoted to a consideration of the scaling of tests *with* and *without* emphasis on a reference population, it is the purpose of this paper to present some of the more common scaling methods and to comment on the role played by the underlying population.

ROLE OF POPULATION IN SCALING TEST SCORES

A familiar frame of reference is provided by the performance of individuals in a single well-defined group on a particular test at a particular time. Two commonly used types of scales have been derived within such a frame of reference. The simplest are ordinal scales such as percentile scores in which the scale number describes relative position in a group. The simplicity of percentile scores is also their limitation: they do not have algebraic utility. The second type are interval scales where an effort has been made to obtain algebraic utility by definition. The T-scores of McCall represent an interval scale where equal units have been defined as equal distances along the abscissa of a postulated normal population frequency distribution.

A second type of frame of reference is provided by the test performance of individuals belonging to well-defined subgroups where the subgroups have a specific relationship to each other within the composite group. Within this frame of reference both ordinal and interval scales have been derived. Initially the basic problem is to obtain ordinally related subgroups such as grades 1 to 9 or age groups from a specified population for the scaling operation. Age scores and grade scores provide ordinal scales which have had wide utility in the elementary grades. Attempts have been made to obtain the merits of an algebraically manipulatable scale by utilizing ordinal relationship of subgroups but introducing restrictions in terms of the shape of the frequency distributions. Efforts to obtain interval scales within such frames of reference have been made by Flanagan in the development of the Scaled Scores (1) of the Cooperative Tests and by the speaker in the development of K-Scores (2). Cooperative Scaled Scores are based on the assumption of overlapping normal distribution of ability groups and K-scores on the assumption that overlapping grade distributions can be represented by Pearson Type III Curves.

The importance of the particular reference population which is used to determine any such scales cannot be overemphasized. A person scoring at the eighty fourth percentile or obtaining a T-score of 60 in an arithmetic test where the score is calculated for a typical seventh grade is obviously not performing equally to one whose standing at the eighty fourth percentile on the same test is calculated for a below-average seventh grade. Likewise a pupil with a vocabulary grade score of 5.2 obtained from a representative sample of fifth graders in, say Mississippi, is certainly not comparable to a pupil making a score of 5.2 based on a national representative sample. The importance of the particular population in determining the fundamental reference point and size of unit is stressed by the originators of both Cooperative Scaled Scores and K-scores. The ratio between the variabilities of overlapping groups in both Scaled Scores and K-scores is a function of the areas cut off in samples of the overlapping groups by the same points in each of the overlapping distributions. Hence this important characteristic of the basic units in each type of scale depends upon the particular sample selected since it is highly probable that overlapping distributions selected from different populations will have different amounts of overlap at points along the scale.

Psychophysical scaling procedures are also sometimes applied to achievement testing. It is to be noted that resulting scales such as sensed difference units which are based on just-noticeable-differences or equally-often-noted-differences are a function not only of the pupils tested but also of the sample of persons making the required judgments.

PROPERTIES OF SCALE DEPENDENT UPON PURPOSE AND DERIVATION

Test scores are used by administrators, teachers and research workers to make comparisons in terms of rank, level of development, growth and trait differences among both individuals and groups. Hence many types of scales have been developed depending upon the intended use. Each is consistent within itself but the properties of the scales are not completely consistent from one type of scale to another. For example a grade scale is not appropriate for measuring growth in a function unless one is willing to accept the assumption that growth is linearly related to grade. K-scores which were designed to provide an interval scale for measuring growth during the elementary school within a particular school subject are not comparable from one school subject to another unless one is willing to assume a common growth for all the subjects being compared. Furthermore the adoption of a uniform standard deviation of 7 K-units for fifth grade distributions defines as equal the variability of fifth grade performance in all functions. The scaling of the Binet items involves the assumption of a linear relationship between Mental Age and Chronological Age. As valuable and useful as the Binet Scale has been for the purpose for which it was designed, it has obvious limitations when we try to infer the "true" nature of intellectual growth.

SCALING STABILITY IN LARGE REPRESENTATIVE POPULATIONS

Scales derive their properties in two ways—by definition and experimental verification. Using K-scores as an example let us consider two desirable properties of a scale: (1) that it shall be invariant with respect to the sample of items used and (2) that it be invariant with respect to the population used in its derivation. The first property is inherent in the definition of K-scores and in the specific definitions of other scores such as Cooperative Scaled Scores. That is, since K-scores are defined by the amount of overlap between adjacent grade distributions any test of a function that will reliably rank the scaling sample in the same way will give rise to exactly the same set of K-scores.

For example, the K-scores obtained from Stanford Achievement Word Meaning test data would be identical to the K-scores obtained from the Metropolitan Vocabulary test data provided all children in the grade range scaled were ranked in the same order by both tests.

The second property mentioned is not necessarily inherent in K-scores in terms of their derivation. With suitable attention to sampling problems it is reasonable to expect to obtain scales with reproducible properties from one sample to another. In contrast such reproducibility is not expected from population to population. There are, however, practical situations in which it would be useful to have a scale which was invariant with respect to more than a single population. For example, since achievement tests are used for measuring growth and comparing the performance of groups over a long period of time (8 to 10 years) it would be desirable to have a scale which would be invariant with respect to national samples taken annually. Any such property of K-scores or any other scale must be established on an empirical or experimental basis. Such stability from one population to another is evidenced in recent efforts to apply K-score scaling to the forthcoming edition of the Stanford Achievement tests.

Grade means, differences in grade means, grade standard deviations and grade skewnesses expressed in K-units determined from the performance of the national normative sample obtained in 1952 on Form J of the forthcoming revision of the Stanford Achievement Test are compared in Table I with the corresponding statistics expressed in K-units determined from the 1940 national normative sample on Form D of the Stanford Achievement Test.

A K-unit is defined as one-seventh the standard deviation of the national grade 5 frequency distribution in any trait where Pearson Type III Curves have been fitted to it and to the adjacent grades in such a way that the proportion of cases in each grade exceeding each raw score is the same as that found in the original data. The mean performance of children in the United States after completing the ninth grade was selected as the reference point and assigned a K-score of 100.

The 1940 sample in terms of which the 1948 K-units were defined consisted of approximately 50,000 cases and was itself a twenty percent random sample selected from about 300,000 pupils to whom the Stanford Achievement Test Form D was administered at the end of the school year in 1940. The sample appeared representative of the national elementary-school population with respect to sex, I.Q., age and geographical location.

The sample in terms of which the present (1952) K-units for arithmetic reasoning were defined consists of approximately 94,000 cases and was selected from a sample of about 460,000 pupils to whom the new Stanford Achievement Test Form J was administered in April and May, 1952. Communities were selected to give a representative national sample in terms of size and geographical location according to the United States census. All pupils in at least three consecutive grades in those communities were tested and a twenty percent sample of these testees was taken at random from each class tested within those communities.

In order to compare the results obtained when K-units in arithmetic reasoning were derived independently from each population let us now examine: (1) the average growth in arithmetic reasoning from grade to grade; (2) the extent to which the variability in arithmetic reasoning changes as children progress through the grades; (3) the effect of progress on the skewness of the grade distributions and (4) whether the fitted curves approximate normal curves.

Although it is commonly believed that growth in specific subjects in the elementary and junior high schools is not constant from grade to grade, the objective verification of this belief has been difficult due to lack of an interval scale extending over the range of grades. The differences in mean achievement (in terms of K-scores) of successive grades of the 1940 sample and the 1952 sample in arithmetic reasoning are given respectively in the fourth and fifth columns of Table I. These differences are indicative approximately of the amount of growth in the trait measured in the particular grade listed.*

The relative change in variability of the performance of children in successive grades has also been difficult to determine, due to the lack of an interval scale extending over the range of grades. The standard deviation in terms of K-units of each grade in arithmetic reasoning for the 1940 sample and the 1952 sample are given in columns six and seven of Table I.

One of the major findings presented in a paper given at the 1948 Invitational Testing Conference was the *consistent increase* in variability in two arithmetic traits from the second grade to the ninth in contrast with two verbal traits in which the standard deviations were *nearly constant.* The present study supports the previous finding concerning increased variability from grade to grade in arithmetic reasoning. The standard deviation in grade 2 is 3.3 K-units, while in grade 9 it has increased to 18.1 K-units. Thus one of the several implications one can draw is that as children progress through the grades the problems of the arithmetic teacher increase in that the groups become more heterogeneous.

The skewness of each grade for each sample is given by columns eight and nine. In the 1948 paper no consistent skewness trends comparable to those observed for grade standard deviations were evidenced. In the present situation there does appear to be an increase in skewness from grade to grade with a single reversal between grades 2 and 3. These data coupled with the previously reported data (3) would lead one to believe that the assumption of normality for every grade distribution is not as tenable a hypothesis as the assumption that the grade distributions are skewed.

---

* However, since the people in each grade were different from those in other grades, these differences in grade means may be considered as growth only to the extent that we are willing to consider, for example, the present third graders as comparable to what the second graders will become a year hence. True growth could be determined by measuring the same people with comparable instruments in terms of K-units at different grade levels as they progress through school.

The data in Table II which were published in the Proceedings of the 1948 Invitational Conference on Testing Problems have been included to show the contrasting results obtained between arithmetic functions (arithmetic reasoning) and a second function (paragraph meaning) when measurements are made in terms of K-scores.

Considering the facts that different tests were used, and also samples from different populations reflecting the lapse of a twelve year period which included World War II with resulting dislocations of pupils and teachers and many curriculum changes, it seems to the author that discrepancies in the *pattern of differences* in grade means and grade variabilities in the two sets of arithmetic reasoning data are minor compared with the general pattern of agreement.

Role of Population in Scaling Individual Items

The problems involved in the scaling of individual test items are similar to those of scaling test scores in that an item may be considered as a test which represents a smaller sample of behavior than the total test score. One of the most widely used scales in which individual items were scaled is the Terman-Merrill scale for the Stanford Binet (4). Items were located on this scale as a result of the performance of well-selected age groups. In his recently developed latent structure analysis Lazarsfeld (5) has presented scaling methods which involve the assumption of a polynomial trace line for each item. The responses of the sample of people to the item are used to determine the parameters necessary to define the scale. In all cases the empirical data which define the scale are dependent upon the reference population used.

In some instances scaling based on total test score is preceded by a scaling or partial scaling of items. In the Stanford Achievement Tests difficulty indices for each item were computed for well-defined and well-described grade groups. A test composed of these items was then administered to a national sample of each grade group and various types of scales based upon the total score were obtained.

GENERAL CONSIDERATIONS

In conclusion this paper has attempted to achieve two objectives (1) to review some of the more common scaling techniques and emphasize the importance of the role of the reference population as a background for the second paper which treats the topic of scaling techniques which minimize the reference population and (2) to illustrate that stable results can be obtained with different large reference populations as shown by an empirical study on the comparability of arithmetic reasoning K-scales based on two national samples of elementary school children taken 12 years apart and obtained from two distinct though similar instruments.

It is to be noted that not only in the argument of this paper but in the development of K-scores (our major illustration) the reference populations have assumed major and fundamental importance. The acceptance of comparable scales utilizing different methods and/or different populations is dependent upon empirical verification.

Situations where there is internal consistency within a number of frames of reference but inconsistency of properties from one frame of reference to another are not unique to scaling. There are excellent analogies in the field of Geometry. The geometries of Euclid, Riemann, and Lobachevsky, each one of which is based on a different postulate about parallel lines, are consistent internally but have certain properties which are inconsistent from one geometry to another. Each of these geometries has its own value and utility as a logical model. The utility of any particular one is determined by the appropriateness or adequacy of the basic postulates to the problem at hand.

One of the objectives of the scientist is to bring together, reconcile and synthesize as many theories and concepts as possible. In the testing field we follow the usual pattern of establishing scales to fit a particular need and then attempt to synthesize the properties of the various scales designed for different purposes. On occasion we find that for complete synthesis we either have to abandon a desirable property or utilize an unacceptable relationship.

Although we continually strive for a single scale with the maximum of desirable properties it would seem inadvisable to abandon useful scales designed for a specific purpose merely because they are not adequate for additional purposes for which they were not designed.

It should be emphasized that the adoption by a test user of any one of the scales available does not exclude the use of any of the others.

In fact, the use of more than one type of scale leads to more adequate interpretation of results in most situations.

TABLE I.

K-Score Means, Standard Deviations and Skewnesses for Each Grade at End of School Year on Stanford Achievement Test Form D Given in 1940 and Form J Given in 1952

Arithmetic Reasoning

Grade	Mean K-Score		Difference		Standard Deviation		Skewness	
	Form D (1940)	Form J (1952)	Form D (1940)	Form J (1952)	Form D (1940)	Form J (1952)	Form D (1940)	Form J (1952)
9	100.0	100.0			11.5	18.1	.73	.86
			2.2	5.3				
8	97.8	94.7			11.2	16.3	.86	.85
			6.5	8.4				
7	91.3	86.3			8.8	11.7	.85	.68
			5.0	7.6				
6	86.3	78.7			8.8	8.8	.38	.45
			5.8	5.2				
5	80.5	73.5			7.0	7.0	.53	.32
			5.2	5.6				
4	75.3	67.9			4.9	6.0	.34	.23
			4.9	5.6				
3	70.4	62.3			4.1	4.2	–.06	–.34
			4.4	4.8				
2	66.0	57.5			3.2	3.3	.16	–.01

TABLE II.

K-Score Means, Standard Deviations and Skewnesses for Each Grade at End of School Year on Stanford Achievement Test Form D Given in 1940

Paragraph Meaning

Grade	Mean K-Score	Difference	Standard Deviation	Skewness
9	100.0		8.0	.29
		3.7		
8	96.3		7.5	.44
		3.4		
7	92.9		7.0	.50
		3.7		
6	89.2		7.2	–.39
		4.1		
5	85.1		7.0	–.10
		4.8		
4	80.3		6.7	.18
		6.5		
3	73.8		7.1	–.40
		6.6		
2	67.2		6.4	.31

REFERENCES

1. Flanagan, John C. *Scaled Scores.* New York: The Cooperative Test Service of The American Council on Education, 1939.
2. Gardner, Eric F. Comments on selected scaling techniques with a description of a new type of scale. *J. Clin. Psychol.,* 1950, *6* 38–42.
3. Gardner, Eric F. "Value of Norms Based on a New Type of Scale Unit." *Proceedings of the 1948 Invitational Conference on Testing Problems,* Educational Testing Service, Princeton, N. J. 1949, p. 67–74.
4. McNemar, Q. *The revision of the Stanford-Binet Scale: An analysis of standardization data.* Boston: Houghton Mifflin, 1942.
5. Lazarsfeld, P. F. (with S. A. Stouffer et al.) Measurement and prediction, Vol. 4 of studies in social psychology in World War II. Princeton: Princeton University Press, 1950, Chapters 10–11.

Selecting Appropriate Score Scales for Tests

LEDYARD R TUCKER

SCALES MINIMIZING THE IMPORTANCE OF REFERENCE GROUPS

SCALES FOR test scores have been the subject of many discussions during the history of mental testing and a variety of procedures attempting to establish scales have been developed. That score scales is still a live topic attests both to its importance and to the absence of a completely satisfactory solution. In light of the extensive literature and the numerous schemes that have been tried for score scales, I view with humility my attempts at contributions to the field. Rather than attempting this morning to present a final, all-encompassing solution, I am going to discuss four propositions which I hope will assist in clarifying thinking about the subject of score scales and then indicate the general nature of several possible procedures.

As indicated by *note 1* on the sheet distributed to you I am limiting my consideration to those situations in which each test yields one numerical score for each examinee and this score is to be interpreted by some person. In effect, I am excluding two classes of tests (1) those tests for which the scores are entered directly into prediction formulas, and (2) those tests for which a number of scores are obtained over the same set of items. When the scores are used directly in prediction formulas, scaling problems for scores on the test are irrelevant to the present discussion. In the case of multiple scores for the same test performance, the situation is more complex than the situations I wish to consider at this time. Some of the propositions and conclusions are likely, however, to carry over to the more complex situation. These two restrictions will not reduce the area of discussion greatly. A point worthy of note is that we have not excluded subtests or sections in a test battery when there is one score per subtest or section.

During this paper I consider it to be axiomatic that the score of a person on a test is used to represent the performance of that individual on the test. The first proposition emphasizes the information given by a test score alone. In the general case, any particular score may arise from any of several test performances. Consider, for example, an eighty-item omnibus test composed of twenty items each of vocabulary, reading comprehension, numerical computations, and figure analogies. A score of sixty items right could be obtained by a number of combinations of items answered correctly. One individual may have answered correctly all items except the figure analogies while another person may have answered correctly all items except the reading comprehension. The score of sixty does not differentiate between these two candidates.

As a general principle I consider it to be obvious that the meanings which may be given to a test score depend not only on interpretations attached to the scores by various studies after the test is constructed but also on the test itself. A test should be conceived in terms of the scores that will result. The kinds of meanings that may be attached to test scores are directly related to the nature of the behavior of examinees the test provokes and to our methods of observation. Proposition I indicates the possibility of ambiguities among meanings that may be attached to a single score. Differentiation among possible meanings of a single score is impossible on the basis of the score alone. The information given by this score is a complex of the possible meanings.

A corollary to Proposition I might be stated that the information transmitted by a particular score would be more definite the more nearly equivalent were the possible meanings of the score. I will return to this point in discussion of Propositions III and IV.

In Proposition II, consideration is given to the significance of differences between two scores. For an example, consider a speeded verbal reasoning test. Score differences in the lower range of scores may be indicative of differences in a complex of verbal comprehension and reasoning abilities. In the higher score range, score differences may be associated to a greater extent with differences in speed of reading. The proposition recognizes not only the possibilities of changes in the nature of differences in test performances associated with score differences at various score levels but also the possibility of changes in the extent of differences in test performances associated with uniform-sized score differences at the various score levels. In some score ranges differences of some given amount between two scores may have much less significance than the same-sized score differences have in other score ranges.

I consider Propositions I and II to be true of all tests no matter how the tests are constructed. Aside from pointing out the indivisible character of any single score, these propositions indicate the existence of a maximum of freedom as to possible meanings of scores and differences between scores. In fact the freedom as to meanings may be so great that a complete loss of significance may occur. The problem is one of being able to limit the possible meanings so as to obtain such definite information as is desired. In an ideal test, as indicated in Proposition III, the information given by each score would imply a single interpretation. Uniform-sized differences between scores would indicate a single kind and extent of differences between corresponding test performances. Such a test would maximize the definiteness of the information transmitted by the scores.

It is to be noted that the establishment of the concept of an ideal test does not limit the nature of the score continuum. In the present state of the art of testing we will probably be able to approach closer to this ideal in some areas of skill and knowledge than in other areas. No matter how poorly or well we can approximate an ideal test for a characteristic we wish to test, the concept of an ideal test indicates a worthwhile goal. The more nearly we can approach this goal of an ideal test, the more definite will be the information given by scores on the test.

Proposition IV emphasizes the point that a unitary continuum may be achieved in a variety of ways. It is important, though, as indicated in note 2, to establish the homogeneity of each continuum considered for an ideal test. Unless the continuum is homogeneous in some sense, an ideal test is impossible. Ambiguities of score interpretation are the natural result of heterogeneity in meanings of test scores. In order to obtain definiteness of information transmitted, it is imperative that measures be taken to obtain homogeneity of the score continuum in some desired sense.

I have listed in Proposition IV two senses in which the score continuum may be homogeneous. The first sense depends on discovery of homogeneous traits in the behavior of the examinees. This is the sense basic to considerable work in psychological research. In contrast, the second sense depends on a homogeneity of evaluations of behavior. One might judge two distinct behaviors of individuals as being of equal value in some field. The fact that the occurance of these behaviors is uncorrelated for a group of examinees would be irrelevant. For example, consider a test such as "understanding of social environment." An understanding of an economic principle such as the law of supply and demand might be valued as highly as understanding of current political events. The homogeneity exists, if at all, in the opinions of the examiners and not necessarily in the behavior of the examinees. It is important to distinguish these two senses in which the score continuum may be homogeneous. Quite different modes of procedure are appropriate in development of tests and score scales for these two senses of score homogeneity.

Turning our attention now to the test development and score scaling problem for each of the two senses in which the score continuum may be homogeneous, consider the first sense, homogeneity of behavior. A number of techniques, including correlational and factorial analyses, have been developed to study the homogeneity of behavior. General agreement exists that the individual differences within a homogeneous domain of behavior will produce a hierarchical table of intercorrelations for the population under consideration. A supplementary type of study involves the difficulties of items as indicated by proportions of groups of examinees who give particular responses. The population for which the test is to be appropriate would be divided into select groups on the basis of whatever available information is relevant to performance on the test. A sample of people in each group would be examined and the difficulties of the items would be obtained for each sample. The sets of item difficulties should be systematically related. In a present experiment, vocabulary test materials were administered to students in the seventh and tenth grades of schools located in each of four categories defined by high versus low socio-economic districts in which the school is located and by location in the north-east versus south-east regions of the United States. One group, thus, includes schools located in low socio-economic districts in the south-east and a second group included schools in high socio-economic districts in the south-east. Two similar groups were defined for the north-east. Item difficulties will be determined for each of these groups. Our question is whether the items will retain the same rank order in difficulty when the item difficulties are based on such different groups. Only in case that both such invariance of rank order in item difficulty and a hierarchical table of correlations exist should the domain defined by the items be considered as homogeneous.

Once homogeneity in a domain of behavior is established, a score scale is to be determined so that each score represents a particular point on the continuum. One might establish groups of items of equal difficulty and arrange these item groups on a difficulty scale. The preceding check on invariance of rank order of item difficulties facilitates this step of grouping items. An examinee would be placed on this scale by the group of items that he was able to perform at a just satisfactory level. On groups of easier items, the examinee would perform better than just satisfactorily and on groups of more difficult items he would perform less than just satisfactorily. The examinee's score would be defined in terms of the group of items he is able to perform at a just satisfactory level. The test user would interpret this score directly as the level of proficiency which these items represent.

An alternative procedure is to use a score on the test composed of the items to establish sub-groups of examinees having approximately equal ability in the function under consideration. All examinees whose scores were within some narrow class interval of scores would constitute each of such sub-groups. The item difficulties would be obtained for each of these sub-groups and a check on invariance of the rank order of item difficulties would be made across the several sub-groups. In case the rank orders were stable, a scale of item difficulties could be established. Each sub-group would be located on this scale by those items with difficulties for that sub-group at some defined level, say 70% correct. It is to be noted that this scale does not depend on the number of examinees who are placed at any particular score value, only the proportions of examinees giving the correct answers to the items are used. The scale is independent of the shape of the frequency distribution of scores. The homogeneity check for the population guarantees independence of the scale from the particular group of examinees used to establish the scale values.

Consider the second sense in which a score continuum might be homogeneous: each score indicating placement on an evaluative scale. Our methodology can now turn to investigations of the opinions of people who will be considering the evaluations. Do the value opinions form a homogeneous field? Or, do separate "schools of thought" occur which would alter the relative order of examinees in the evaluations given by members of different schools of thought? Methodological developments are in progress in the field of psychometric scaling methods which show promise for application to this problem. Once homogeneity of opinions is established for some defined domain and a scale of values for behavior is developed, a test may be constructed which will locate individuals on this scale of values. Points on the scale would be defined by those behaviors which were values at those points. Such scales would depend directly on the group of judges making the evaluations and would depend only indirectly on the behavior of the population to be examined. A question could still remain as to how opinions are influenced by observations of behavior.

When either type of scale indicated in the foregoing discussion has been established, a survey of performances of a population would be desirable. Comparative data between examinees would then be produced. The advantage of use of ideal tests, as here conceived, and the resulting scales in the survey operation is that definite information about the distribution of the population as to levels of performance would be obtained.

Note 1: Consideration is limited to those situations in which each test yields one numerical score for each examinee and this score is to be interpreted by some person.

Proposition I: Each test score by itself transmits the same information for all examinees who receive that score. This information may indicate some complex of qualitative and quantitative characteristics of the test performances.

Proposition II: Considering for one test two raw scores differing by one unit, the information transmitted indicates some complex of qualitative and quantitative differences between the two test performances with the kind and extent of these differences in test performance possibly changing from one score level to another.

Proposition III: An ideal test may be conceived as one for which the information transmitted by each of the possible scaled scores represents a location on some unitary continuum so that uniform differences between scaled scores correspond to uniform differences between test performances for all score levels.

Proposition IV: The score continuum for an ideal test may be homogeneous in any of a number of senses. Two basic senses are:

1. The scores indicate extent or degree of some trait which exhibits homogeneity in the behavior of examinees.
2. The scores indicate placement on an evaluative scale for a category of behavior considered in a unitary fashion by those people making the evaluation.

Note 2: It is important to investigate the homogeneity existing for the sense in which the scores are to form a continuum.

Note 3: For each of the two senses listed in Proposition IV, experimental and analytic methods for test development and score scaling may exist or be developed which do not depend on the relative number of examinees who receive each particular score in a reference group of examinees. Such methods would yield scaled scores indicative of levels of performance on the test rather than a comparison of relative positions of examinees in a group. The comparisons of examinees may be performed as a separate, later step.

Selecting Appropriate Score Scales for Tests

JOHN C. FLANAGAN

DISCUSSION

In discussing the matter of how we were going to divide up the discussion, Dr. Lindquist and I did not have a chance to review each other's remarks, partly because they were not entirely formulated, there was some delay in receiving the remarks, and partly because we thought it might interfere with the spontaneity of the discussion.

We agreed that we would choose different topics. He is to talk about the basic considerations involved in this problem and I am discussing the fundamental principles.

One of the fundamental principles we have to deal with in this problem of scaling of test scores is that we do not have any ideal scores such as the ones that Ledyard Tucker has been talking about. In practically all cases of tests with which I am familiar you have much more information if you know exactly what the response of each man to each item is than if you have a simple summary score. In other words, we must remember we do not have ideal tests. Presumably Dr. Tucker is talking about an ideal mathematical model which we will never have in practice. We may approximate it in many situations but for practical purposes it is just something to think about, not something that we will be able to use.

For example, if we take a test of history, it is ridiculous to assume that one teacher's group will not learn more about some particular types of items concerning Betsy Ross or the Civil War or some other happening than some other teacher's class. Therefore we will never have this perfect homogeneity which is necessary in most types of achievement tests. The one place where we might get something approaching homogeneity is in some sort of power scale. We have a power scale in situations in which the items can be so arranged that if you can do a specific item at one point on the scale, you can obviously do all those earlier on the scale. In such scales we do not have any specifics. Training affects all items equally. Nothing that happened yesterday morning or that you read in the newspaper can affect one item and not the others. When all these conditions are fulfilled we have a homogeneous scale.

It seems unlikely that an educational test of this type will be found and therefore it appears that this is a model to think about and not something for practical use. Certainly such a test will not be found of vocabulary. Obviously it is silly to think about a perfect order of difficulty of vocabulary items. They are going to differ in difficulty according to what you have looked up in the dictionary recently, or what somebody else has popularized. Similarly for mechanical principles items. Specific experience is going to prevent you from ever having one of those homogeneous behavior scales.

The other type of ideal scale suggested by Dr. Tucker involving homogeneity of opinion evaluations seems even more remote from any possibility of realization. We usually use as our score the number of right answers, or some function of this, recognizing that this is an oversimplification. We would theoretically be better off to weight some items differently. Each of these items is not of equal value and equal importance; it cannot be exactly as important as each of the other items in determining what we are trying to determine. It must be recognized that having the exact pattern of how each person performed on each item contains more information than we can get out of any single score.

Assuming that we have a simplified abstraction of this performance in terms of a score, what is the fundamental principle for determining how we should express these scores? It seems to me that the fundamental principle governing our behavior here is utility, what scores are going to be most useful to us. This depends on purposes. I would think that there certainly are, as Dr. Tucker said, some purposes for which certain test scores will be more valuable than others.

For many types of analysis and study of test scores such as for prediction problems, we are going to want to use the scores to calculate product moment correlation coefficients. It is very desirable in getting an estimate of the correlation in the particular population to have the same shape of distribution for the scores of the two variables. In other words, if you have a skewed distribution in one variable, you want a distribution with the same skewness in the other variable. This will yield the maximum possible correlation between these two variables. It seems to me, therefore, for a lot of purposes if we can make the distributions normal, we have more likelihood of obtaining consistent results than if we skew one of them one way and another one another way, and so on.

On the other hand, if we normalize them from one population and find that they are radically skewed for others, this perhaps would suggest that some variation in the basic type of scale should be introduced.

I should like to review the fundamental principles in establishing a set of scaled scores. The first of these is the reference point. Assuming that you are going to modify your raw scores, there is no use changing by adding 10 or 15 or subtracting 20 or something from your raw scores or making some other change, unless you are going to get some meaning into this fundamental point of reference.

One of the most useful scores that we have had has been the I.Q., because of the simple meaning of a score of 100. The I.Q. score has other difficulties, as pointed out in meetings here in the past couple of days, but the fundamental point of reference of 100 has been extremely useful. We tried to capitalize on this type of thing in getting a fundamental point of reference for scaled scores. The 50-point is very similar to 100 I.Q. in its fundamental meaning. Similarly, in establishing the stanine scale, we have established 5 as a point of reference.

The second problem is the size of unit. It is important that this size of unit have some meaning. Some simple, easily remembered meaning is desirable. Making the standard deviation equal to 10 or 2 or some simple multiple of this sort is useful because many of us dealing with such scores remember the unit normal distribution and therefore know about how many scores can be expected to be as far as two standard deviations from the mean. This tells us immediately something about the scores.

The other question involved in size of unit is coarseness. In the scaled scores we used a standard deviation of 10; in stanine's, we have a standard deviation of 2. There are some scales using standard deviations of 100. These are 3 digit scores as compared with the one digit that have been used frequently. I doubt that there are very many tests which justify three digit scores because of their accuracy of measurement and the uses to which they will be put. In the military services with a day and a half of testing and 10 or more scores going into each composite, we still reported the composite on a 9 point scale. Certainly to put a 15 minute or a half hour test on a 3 digit 1000 point scale seems a little ridiculous.

In some new work which I am doing on aptitude tests, I have decided that a 27 point scale in which you used the 9 point scale with a plus, minus and zero would be a little preferable to the 9 point scale. For most purposes people are not going to pay much attention to the plus, minus and zero, but for some purposes, especially where you are dealing with a group, all of whom are at the high end of the scale, this might have some value in breaking ties, although reliability of the scores may not be sufficient to make this breaking of ties of very much practical advantage to the user.

The question of equality of units is the last of the fundamental principles to be discussed. The problem of utility is the primary factor here. We want to get distributions which have as similar shape to other distributions as possible and with as much constency of shape or distribution as possible from one population to another. I think this involves a certain amount of trial and error. If we find that one particular shape, say the rectangular distribution as from percentiles in the sixth grade would provide basic units which would distribute results from the fifth and seventh grades rectangularly also and show similar consistency from one region of the country to another, I would certainly say we ought to use units yielding rectangular distributions. I think, as most of us have experienced, this does not happen. We are much more likely to get consistency if we have a normalized set of scores as the basic units.

As to whether or not some element of skewness is important for some situation, I think we do not have adequate information yet. Certainly this field should continue to be explored along these lines Dr. Gardner has been following.

One other point that should be made is that in setting up these scales we should distinguish between ideal properties of scaled scores and practical factors of convenience and cost. In planning for the battery of aptitude tests which I am publishing shortly, a comprehensive review of all the circumstances suggested the best thing to do would be to have the stanine of 5 represent a random sample of 18-year-olds in the United States population. Having thought about this and having explored the possibilities of getting this done through draft boards and similar means, I finally rejected it. I wish to make it very clear that it still seems the best thing to do but not within my resources. As a practical expedient we are adopting as the basic reference group Pittsburgh public high school seniors.

Certainly such a reference group is homogeneous and has certain advantages. However, we should make it very clear in our discussions whether we are talking about what we ideally think we ought to have done. I still think that a random sample of 18-year-olds would be better. That just did not seem to be feasible to me at the present time.

I think, in closing, that we should try to keep in mind that any system of developing a series of scores should be for practical utility and it should be demonstrably more useful to people who are going to use and interpret the scores than the other procedures, raw scores that are available to them.

Selecting Appropriate Score Scales for Tests

E. F. LINDQUIST

DISCUSSION

THANK YOU, John, for sticking so close to the agreement. This has worked out pretty well. I can see almost no overlap or similarity between John's fundamental principles and my basic considerations.

What I propose to do is to present to you a number of what I have termed basic considerations in scaling educational achievement tests. These are designed to support a particular point of view with reference to the whole problem of scaling. You might say that their purpose, so far as this audience is concerned, is to create an attitude or a change in attitude, if possible, toward the scaling problem.

I think it might be well to start off by specifying some of the purposes of scaling. Perhaps this is so obvious it hardly needs to be said, but I would suggest that scaled scores are needed for three general purposes; first, to facilitate comparisons between performances on different tests and thereby to provide a basis for computing properly weighted composites of scores on different tests. Second, to facilitate comparisons between differences in performance at different levels for the same or different tests; and, third, to facilitate the presentation of normative data, either by incorporating some of the normative data in the scaled scores themselves, or by making it easier to organize and present the tables of norms. It is much easier to prepare manuals or tables of norms if all can be referred to a single reference scale than if one has to refer each to a raw scale for each individual test.

Of these three purposes, perhaps the most important is the first although I am not interested now in arguing the relative importance of these purposes.

Perhaps before going further, I ought to say also that my remarks will be pretty much restricted to applications to educational achievement tests, and will not include psychological tests of various types, aptitude tests, interest inventories, and that sort of thing. I should like to begin, then, by another obvious statement, defining what I mean by an educational achievement test. An educational achievement test is one designed to reveal differences among the examinees in the extent to which they have attained a particular educational objective, or set of objectives. With reference to that definition, a good educational test, it seems to me, must by itself constitute a complete and adequate definition of the objective with which it is concerned. That is because, as it works out in practice, the test itself so frequently becomes the end of instruction. Teachers and principals make it their business to improve the average score on many of these tests, and unless the things that they must do in order to achieve higher score averages are precisely the things that we would like to have them do —in other words, unless the things measured by the test are precisely the educational objectives to be achieved—we are going to get into serious difficulty.

A good educational achievement test, then, must itself define the objective measured. This means that the method of scaling an educational achievement test should not be permitted to determine the content of the test or to alter the definition of objectives implied in the test. From the point of view of the tester, the definition of the objective is sacrosanct; he has no business monkeying with that definition. The objective is handed down to him by those agents of society who are responsible for decisions concerning educational objectives, and what the test constructor must do is to attempt to incorporate that definition as clearly and as exactly as possible in the examination that he builds.

Now, the statistical properties of educational achievement tests and of test items are to a very large degree a function of arbitrary features of the school curriculum and of variable features of the examinees. Dr. Gardner gave some very convincing evidence of that. I should like to add just a little bit to it only by way of samples of the kind of thing I mean. I have the data on a few arithmetic test items that were tried out on a very large population of elementary schools in the Iowa Basic Skills Testing Programs. For instance, one of these items reads, "multiply 506 by 8." The difficulty of that item in the third grade was 4 per cent—that is, 4 per cent answered that item correctly; in the fourth grade, 55; in the fifth grade, 84, and from there on it maintained that high level.

Another item reads, "divide 84 by 2." At the beginning of the third grade this item has a difficulty of 13 per cent. By the end of the second semester of the third grade it has a difficulty of 57 per cent. By the beginning of the first semester of the fourth grade it has a difficulty of 83 per cent.

One more item: "add ⅓ and ⅗." This has a difficulty below 5 per cent in each of grades 3 and 4, in grade 5 it has a difficulty of 5 per cent, in the sixth grade the difficulty jumps from 5 to 78, and in the next grade it is 86.

Now, what accounts for these abrupt changes in the difficulty of the item? Is it attributable to some natural change in the nervous or physiological maturity of the child? Obviously not. It depends only upon certain arbitrary decisions that have been made in the organization of the school curriculum. The schools have decided to teach this item in this grade, and that item in another. Tomorrow they might change their minds. There is nothing magic about these abrupt changes.

Those changes in difficulty occur not only from grade to grade, but even from school to school within the same grade. The item, "divide 84 by 2," was tried out in 12 different schools the same time of the year and under the same conditions. These schools together yielded a sample of about 600 pupils. In School A, the difficulty of the item was 39 per cent. In School B, it was 6 per cent. In School C, it was 82 per cent. In School D it was 100 per cent. From School B to School D, the range in difficulty of this one item was from 6 to 100 per cent.

The item, "subtract ⅚ from ⅔" has a zero difficulty in School A; an 85 per cent difficulty in School B.

The item, "multiply 0.24 by 52.4" has a zero difficulty in School A, and 82 per cent in School B, and a 0 per cent difficulty in School D. Clearly, this is because these schools could not agree upon the point at which this item was to be presented in the curriculum, and so in one school the item was extremely difficult and in another school it was very easy.

The decisions that characterize the differences between these schools could easily characterize entire populations. All of the schools in one population might decide to teach one item in one grade level and all of them in another population to teach the same item at another grade level.

All methods of scaling educational achievement tests now being considered are based upon the statistical properties of the test or of the individual items constituting the test with reference to a particular population of examinees, which was Dr. Gardner's main point, and with which I would agree absolutely. That is, all scales are derived from normative data.

Now, raw scores on some educational achievement tests are meaningful in themselves in terms of the content of the test. For example, you might build up a test of the 100 basic, addition combinations in arithmetic, and you might find that a particular student answers correctly 60 out of those 100 items. That obviously means something without regard to anyone else's performance on that test. This is what we might, for purposes of this discussion, call an *absolute* meaning, or a fundamental meaning. But groups of items arranged with reference to such meanings do not constitute scales. You cannot compare *60* out of 100 basic addition facts with *5* out of 20 rules of grammar, or with a certain number out of a possible number of vocabulary items in French, and so on. Even though you have grouped the items in this respect, you must still attach numbers to those groups that will make the performance comparable from group to group. In other words, the scaling job still has to be done after this grouping has taken place.

Any meaning that a scaled score has, in addition to that contained in the raw score, it has because of the normative data incorporated in the score, and that meaning applies strictly only to the particular reference population involved in the scaling process. In other words, no scaled score has any fundamental meaning attributable to the scale itself. Whatever meaning it has, in addition to the kind of meaning I just discussed, it has because of the normative data incorporated in the score.

It is impossible to incorporate in any single scale normative data for more than one reference population. However, in order to interpret satisfactorily the scores on most educational achievement tests, one must refer to data for a large number of different reference populations. The kind of tests we are talking about, elementary school achievement tests, make that very obvious. You have a different distribution of scores on the test for every one of the grades, say, from the third to the eighth. You have another different distribution of scores if you throw all of those grade populations together, which is the population that we use in effect when we establish a grade equivalent scale. Again, you may wish to interpret school averages, and they must be interpreted with reference to distributions of school averages, not with reference to distributions of pupil scores—and there again you have a different distribution for every grade, and you have a different distribution for all grades thrown together.

Furthermore, as has been suggested, you will get different distributions within the same grade for the same test from one geographic population to another. There are marked differences in Iowa between the distributions of scores for one-room rural schools and the distributions of scores for schools in communities of more than 25,000 in population, on exactly the same test. It is thus possible to identify a very large number of reference populations, all of which must be used for satisfactory interpretation of the scores on any of these tests.

Accordingly, whatever we choose to regard as the basic scale for a test, we must always set up alongside this scale a large number of other scales, or if you prefer to call them that, tables of norms, each of which is to be employed for a different purpose. In most, if not nearly all, practical situations it is difficult to determine which of these purposes is most important, or which scale is really the basic scale. Indeed since the relationships among the various scales can always be determined, that is, with reference to a particular population it can always be determined, or since any scale can always be expressed in terms of any other scale for that population, there seems to be very little point in trying to determine which scale is basic. However, it is usually desirable for reasons of convenience to select one scale to be employed as a reference scale. This is the scale to which the raw scores are often immediately converted in the scoring process, and in terms of which the only original record of the scores is made.

With reference to this reference scale, from at least one point of view—perhaps I should say before presenting this point of view it is not my own, it is not one by which I would abide in practice, but it certainly is a point of view that deserves consideration—the best type of reference scale for a test is one that is divorced as much as is possible from any normative meaning. I repeat, from one point of view, the best kind of reference scale is one completely devoid of normative meaning. For example, the scaled scores along the reference scale for a test might be simply the corresponding raw score expressed as a per cent of the possible score on the test. That might have some absolute meaning of the kind I discussed earlier, but it would have no normative meaning. It would not be a good reference scale, I want to say at once, because there are too many connotations, undesirable connotations attached to the per cent score on the test. But the use of a scale that is divorced from normative meanings has the very distinct advantage that if the norms change after the scale has been established—and that does frequently happen—then there is no need to abandon the scale on that account, or to rescale the test. Instead, all one need do in that case is to leave the reference scale as it was before, because it does not depend upon normative meanings, and make whatever changes in the normative scales associated with it happen to be appropriate, and those changes may affect only some of the normative scales.

Let me now attempt to summarize what I have been trying to say. For educational achievement tests—I mean specifically tests like the Stanford and Metropolitan Achievement test batteries, the Iowa Tests, Basic Skills and the projected ETS tests of basic educational objectives—the amount of meaning that can be built into any single reference scale will constitute only a very small part of the total amount of meaning to be derived by all of the test users from those test results. To a very considerable extent, therefore, any so-called basic scale is primarily a device for facilitating the presentation of other scales. The scale value of a given performance on any of these other scales will be exactly the same for a given reference population regardless of the nature of the reference scale used, because there is always a monotonic relationship among these scales. Accordingly, the problem of what size scale, or what kind of a reference scale is to be employed with an educational achievement test is a problem, in my opinion, of relatively minor importance. The major problem is what *scales* should be employed with educational achievement tests, scales in the plural rather than scale in the singular. That is, what kinds of norms should be provided with the test, and how and for what purposes each should be interpreted.

I should like to conclude with just one or two more specific comments with reference to the proposals or suggestions that have already been made. So far as these so-called basic considerations are concerned, I would certainly have no objection to the use of a scale such as Dr. Gardner suggests. I would only want to point out you might have to establish a scale of that kind for each of a very large number of different reference populations, because a scale of that kind established for one reference population will not serve all of the purposes that have to be served, or even a very large proportion of them.

I should like to point out, also, that comparisons were made by Dr. Gardner of score distributions on the K scale for different grades. Now, with the K scale certain assumptions underlying those comparisons. The first is that there is a common growth curve for all of the tests involved, but what kind of a growth curve you have for a particular test depends, as pointed out a moment ago, upon what arbitrary decisions have been made by the schools with regard to the presentation of those particular skills or abilities in the curriculum.

The K scale also assumes uniform within grade variability, but what the relative variability within a grade for a particular test is depends again upon these arbitrary decisions. We know, for example, that for sixth grade the variability within grades for arithmetic fundamentals, in terms of differences between successive grade medians, is very much smaller in arithmetic than it is in reading. In the case of the Iowa Tests of Basic Skills, the seventh grade norm lies at the 95th percentile of the sixth grade distribution on the arithmetic fundamentals test, but for reading comprehension, the seventh grade median lies at the 65th percentile in the sixth grade distribution, and for a test of basic social concepts which we recently have tried out in Iowa, the seventh grade median lies at about the 55th percentile in the sixth grade distribution. This is a terrific difference in overlap from grade to grade, a terrific difference in relative variability from grade to grade, which, it seems to me, is obviously attributable to differences in curriculum decisions with reference to grade placement.

Now, finally, with regard to Dr. Tucker's proposal, I will say very much the same thing that Dr. Flanagan has said. If Dr. Tucker does succeed in finding a number of items that all happen to have the same rank order of difficulty for a particular group of reference population, he will find those items only because it *happens* to be true of the school curriculum that the schools have attained some agreement on those particular items, and that will be more or less an accident rather than anything fundamentally descriptive of the child.

Furthermore, if he does find a number of items of that character, those items will *not define* any of the educational objectives which have been handed down by such committees as the Mid-Century Committee on Educational Objectives. It will be a matter of accident what those items define in the way of an educational objective. So I would say that while I would be perfectly willing to accept the kind of scale that Dr. Tucker suggests, I would not—because of the second principle that I suggested, that a good achievement test must by itself constitute an adequate definition of the objective with which it is concerned—on that account, accept the kind of scale that Dr. Tucker suggests.

ADDRESS

DAVID C. McCLELLAND

THE MEASUREMENT OF HUMAN MOTIVATION: AN EXPERIMENTAL APPROACH

WHAT I HAVE to say this morning will be somewhat of a change of pace from what you have been listening to, since I approach the measurement problem from an experimental point of view rather than from the traditional testing point of view.

I should like to review first the different ways in which psychologists have attempted to measure motivation in the past. In the first place, the simplest way, apparently, to measure human motivation is to ask a subject how motivated he is for something or other. The psychologist always starts with the simplest approach: just ask the subject. Of course, we psychologists did this and we did it elaborately. We did it by setting up self-rating scales; we drew graphs to shows the normal distribution, and we urged the subject to follow the normal distribution or put his check marks in some kind of a pattern, but fundamentally the method involves simply asking the subject how motivated he is.

I do not need to tell you, I think, what the difficulties with this approach are. One of the major ones is, of course, that subjects have different subjective standards, and if you ask them how motivated they are for achievement, each one will have a different idea of what intense achievement motivation is, so that when you try to compare their self-judgments, you get "hash."

The second approach is, if you can't ask the subject, then ask somebody else how motivated he is. Of course, you choose somebody that knows him fairly well, presumably, like a teacher, and you get the teacher to rate the pupil on how motivated the pupil is. If you don't think teachers are capable of doing this correctly or validly, you can ask a clinical psychologist, who may study the person for several weeks, or even several years if he is a psychoanalyst, and then get him to make a rating as to how motivated the person is for achievement.

And I suppose if the clinical psychologist does not feel he is capable of doing it, you can always ask a psychiatrist, whose judgment may be even better.

But again I do not need to tell you the difficulties with this methodological approach. One difficulty is circumvented: if you ask the same judge to judge several people, he can more or less keep his standards the same, so that you do not have the problem of shifting norms quite as badly as you do if you ask the subjects to rate themselves. But other difficulties arise. For example, it is not exactly clear what the judge is judging, just what his definition of the motive is. His definition isn't always communicable.

The chief objection to this approach is partly practical and partly theoretical. Practically, I do not think that judgments of motivation have proven extremely fruitful in predicting performance, and I suspect that the reason is that the judgments are not pure enough. Too many factors are taken into account in a clinical judgment so that it is difficult to tease out precise relationships with performance. This difficulty ties in with the theoretical objection that I have—an objection which can be highlighted by comparing the process to asking a group of physicists to measure temperature by pooling their judgments as to how hot it is or how cold it is. You can undoubtedly get a reliable estimate, that is, you can get an agreement this way, but it isn't exactly measurement. I have always been interested in pushing our measurement of motivation more in the objective direction.

A third way of measuring motivation, at least achievement motivation, which I am chiefly concerned with here this morning, is to look at behavior—this darling of American pyschology behavior. It is what the person does that counts. It is not what he thinks, feels, or believes; it is what he does, and if he works hard, he has a high achievement motive. Why not use that as a simple method of measuring motivation: how hard does the pupil work? Well, again there are difficulties here. Theoretical psychologists tell us that performance is determined by more factors than just motivation, so that if you use performance as an index of motivation, you get a lot of other things mixed in there, too, such as past learning, intelligence, etc. Another difficulty, even more serious for motivational theory, is this: a person may work hard for several different reasons. He may work hard because he is anxious or worried, not because he has a high achievement motive. So performance can never prove a very adequate method of measuring achievement motivation, per se.

TESTING PROBLEMS

With this background, I want to introduce the method of measuring motivation that we have adopted, I think more or less by accident. I must tell you that a lot of things become clearer in the cold clear light of hindsight than they are at the time. I assure you that five years ago when we began our research on the achievement motive, we did not go through this step-by-step analysis of other measurement methods, reject them, and then choose the one I am going to describe to you. It happened much more accidentally than that, but now that we have done it this way, our line of reasoning looks sensible to me.

What we did was to do *content analyses of imaginative behavior or fantasy.* (1) Why did we choose fantasy or imaginative behavior? I use "fantasy" for my clinical friends, and "imaginative behavior" as a kind of bow to my Yale background. They mean the same thing.

I think our primary reason for choosing fantasy was that it has so obviously worked. Psychologists have had a long history in the clinical field in which free association, fantasy, and dream analysis in the hands of the psychoanalysts have led to very fruitful and productive motivational analyses. If you stand off and look at the whole psychoanalytic tradition, beginning with Freud, you can oversimplify it by saying that it really deals primarily with motivation. Freud was not particularly interested in learning or problem-solving in the great American tradition; he was much more interested in motivation, and I think the reason is partly methodological. What Freud studied was not problem-solving, not learning, not how you get a pencil through a maze; instead, he studied fantasy—free association—and because he studied this type of behavior, I think he arrived at a motivational type of theory.

So we took this as a lead, and, of course, we had the support of the long Murray tradition at Harvard, which had shown some very fruitful motivational analyses based on fantasy.

One might discuss here why fantasy should provide a good index of motivation, but I will not try to do it; it would lead me too far afield. The sort of argument you make is that fantasy is not influenced much by factual statements, by knowledge. It is not much influenced by values, what a person ought to say in a test, as he doesn't have a very clear idea of what he ought to write in a Thematic Apperception Test. So, the reasoning runs, the only thing that is left to determine his responses is motivation, by a process of exclusion. At any rate, we use fantasy for whatever reason.

(2) Why content analysis? Well, content analysis, for my money, is

just a more systematic way of making a judgment. The usual way of treating a Thematic Apperception Test record is to have a judge read the whole record and synthesize his impression of it into a rating. Well, from my comments earlier about the complexities of such ratings and what goes into them, you can see that I would want to move in the direction of a more objective nose-counting operation. A good analogy which I have often had clearly in mind is the process of making blood counts such as a medical technician makes: you get a sample of blood under standard conditions, you put it under a microscope with a grid over it, you count the number of red corpuscles and white corpuscles, etc.

Our approach is somewhat similar: you get a series of thought samples, or samples of imaginative behavior, and then develop a categorizing or classifying system. Then you count the number of times that a certain imaginative element appears. The operation is a simple yes-no dichotomous type of thing, presence, absence—the imagery is either there or isn't there, like the white corpuscle.

Well, so much for background. Now a little more in detail about the procedure. We need three things: first, we need a method of collecting thought samples. Here we modified the Murray TAT technique by obtaining brief written stories from subjects under group testing conditions. In this way we can test as many people at once as you can get into a room in clear view of the screen on which we project the pictures, in response to which the subjects write their stories. So it is a group testing procedure. We put a short time limit on the story, because we didn't want to give people extra credit for verbal fluency. Since some people obviously can write long stories and others can only write very short stories, we limit the amount of time to around five minutes. In this time we obtain a kind of standardized thought sample averaging around 90 words in length.

Secondly, we need a method of scoring for the achievement motive. I will refer to the achievement motive in the Murray tradition as the need for achievement or more briefly on achievement. We need several things as prerequisites for a scoring system. First of all, we need a criterion of achievement imagery; we have to recognize the white blood corpuscle when we see it, so to speak; we have to recognize the achievement imagery when it is there. We developed by a method which I will describe a little later, a scoring criterion which can be briefly summarized in this phrase, a kind of catch phrase that we use: "competition with a standard of excellence." Examples of it can, of course, be multiplied. A person wants to do a good job; he wants to beat somebody else. These are the two types of standards of excellence that you find, the same standards that golfers use in match and medal play. In match play you try to beat the other guy; in medal play you try to beat par. Both these types of standards are included under our scoring criterion of "competition with a standard of excellence."

Next, we need a set of related categories. Having got the imagery criterion, we must be able to identify other thought elements relating to this central category. Here we tried very hard to get a *related* set of categories that had some theoretical sense, that hung together. To do this we simply followed the standard description of the problem-solving behavior sequence that you find in any elementary text book, e.g., the process of adjustment. It is usually represented with an arrow for the motive, with a rectangle for the obstacle which the person goes around to get to the goal (represented again by a "detour" arrow), etc. We defined subcategories for each part of this behavior sequence. I will not go into more detail here, because I assume that you are not interested in the detailed definition of these subcategories.

Thirdly, we wanted a method of scoring that was, as I said earlier, as operational as possible, as simple as possible, so that it could be readily communicated and readily used by scorers. Here, of course, the ultimate test is scorer reliability, the ease with which you get high agreement co-efficients between two trained scorers. We succeeded pretty well. Our agreement coefficients ran around .90–.95 for scorers judging on different occasions if they were well trained. Training, incidentally, takes a week for some people, longer for others. There seems to be an ability factor involved in ease of learning to score such records. If somebody can tell me what it is, I would appreciate it.

Fourthly, we need a scoring system that is as economical and simple to apply as possible. You may ask at this point why we didn't use a multiple choice system so that a machine could do the scoring instead of a human being. It is obviously much more expensive to use a human being and much more tedious, when you have got hundreds and hundreds of records to score. The answer is, of course, that we would like to use a multiple choice test but it doesn't work, and if any of you want to go out and try it, all I can say is, more power to you. We have tried it and it has just never worked. The same seems to be true of the multiple-choice Rorschach. Some day we will know why multiple-choice projective tests don't work. Now there is just plenty of practical evidence that they don't. It may be because multiple-choice introduces a reality factor which tends to minimize the importance of motivational determinants of perception (see a recent experiment by Crutchfield and Postman in the *American Journal*, "Psychology on the effects of hunger on perception").

In any case our scoring system is not so terribly inefficient and uneconomical. It turns out that a trained scorer, if you can keep him at it—which is another problem—can score 50 to 60 records a day without straining himself, and this means if you have ten scorers, you can score five hundred a day. It is practical, in other words. It takes about a minute to score a story, or five minutes to score an individual record, which isn't excessive.

So far we had a method of collecting the thought samples, a method of scoring them, and next we needed a method of arousing the achievement motive experimentally. Here is where we took a new step in the testing field, I believe. That is, we argued that we did not want to have an *a priori* scoring system. We wanted one that reflected sensitively experimentally-induced changes in achievement motivation.

So we began with two groups of subjects: roughly, a control group and a group in which the achievement motive was aroused. Our method was to compare the imagery in the stories written under neutral conditions with the imagery in the stories written under aroused conditions. We found shifts in achievement imagery. Students wrote different kinds of stories under these two conditions, and we used these differences to arrive at the definition of achievement imagery which I gave you earlier. Note the importance of the experimental variable in arriving at our scoring system. We used only those imagery categories which increased in frequency when the motive was aroused. In fact we redefined our categories so as to capture as best as we could the differences in stories written under "control" and "arousal" conditions.

Now for the payoff, if any. You may well ask: all right, you have demonstrated that imagery in stories changes when you arouse achievement motivation, how can you use this to measure individual differences in achievement motivation?

We took several steps here to see whether we were able to measure individual differences. First, we did a very simple and elementary thing. Having decided on our scoring system based on the categories which increased when the motive was aroused, we simply summed these characteristics in a given person's record. Suppose a person writes eight stories: we went through and scored each story separately, according to the achievement motive scoring system, and then counted the different types of achievement imagery which appeared in his eight stories and got his total score. People varied enough in this total score to give us a reasonable spread and we could begin to relate individual differences as measured in this way to other types of behavior. The basic assumption is that if a person shows a lot of the kind of achievement imagery which appears when the motive is aroused, he must have a strong achievement motive.

To what other types of behavior did we relate our achievement score? First and foremost, as you might expect, we were interested in knowing whether the achievement motive, if measured in this way, was related to performance. That I suppose would be the first question any of you would ask: do the students with high achievement motivation work any harder? It seems logical that they should. Our first experiments in this field were done with laboratory tests. If you take a simple test like adding two place numbers and give college students a ten-minute repetitive test of this sort, you find that there is a very significant difference between subjects with high achievement motivation and those with low achievement motivation. That is, we found that the ones with high motivation had a higher output of arithmetic problems; they completed more of them in the time allowed.

Secondly, if you take a more complex task, like unscrambling words, which is a relatively unfamiliar task as compared with adding two place numbers, you find that while the people with high and low achievement motivation start out at about the same output level the ones with low motivation do not improve during a 20-minute test period. Those with high motivation do improve, so that at the end of the test period they are turning out more work per unit time than they did at the beginning. In other words, they are sufficiently motivated to learn new and better ways of unscrambling words.

I suspect that some of you will be interested in whether or not this measure of motivation is related to grades. I am going to leave that until last, because it is a complex question, and treat it separately.

Let me go on first to other types of behavior in the laboratory to which this measure of motivation is related. Take memory, for example. For years the problem of the better memory of incompleted tasks, the so-called Zeigarnik effect, has been something of a puzzle, at least to some psychologists. Why are incompleted tasks remembered better? There have been, as you know, some conflicting results. Sometimes you find this effect and sometimes you don't. We found that one of the variables correlated with better memory for incompleted tasks is achievement motivation. Subjects with high achievement motivation have a better memory for incompleted tasks. Subjects with low achievement motivation generally have a better memory for completed tasks. They remember their successes, as it were. They are a little bit defensive about this. The ones with high motivation, on the other hand, apparently regard the incompleted task as a challenge. They want to recall it so that they can complete it. They think to themselves, so to speak, "If I had only had time to finish that. If that guy hadn't interrupted me, I would have finished it."

Or take level of aspiration—something that you would think motivation should be related to. Here again we found a relationship, if you rule out reality factors. That is, level of aspiration, as most of us have assumed from the beginning, is partly determined by wish factors and partly determined by reality factors. If you ask a person what kind of a grade he expects to get in a course, he will be determined partly by his past performance, by his previous grades in this course, and also presumably partly by his need for achievement.

We found if you just correlate the achievement motive score with level of aspiration, you don't get any correlation, but if you do it when the reality factors are minimized, or are in conflict, when the subject doesn't really have any basis for saying in reality what he will do on a certain test, then you get a very significant correlation with achievement motivation. This, of course, is exactly what you would expect.

Take perception. We have done experiments on the recognition of words with the tachistoscope, and we find, as one would expect, a certain selective sensitivity. The ones with high achievement motivation recognize words relating to achievement more rapidly.

Let me just mention two others. I could mention a great many more, and perhaps if there is time for a question period, you can ask me about them then.

A very popular test nowadays is the F scale, a measure of Authoritarianism. Roger Brown at Michigan tested to see whether achievement motivation was related to the F scale. I must say I did not expect any relationship, but to my surprise, he found one but it was inverse. That is, students with lower achievement motivation are generally higher on the authoritarianism scale. I think you will see why this may be so in just a minute.

Some of you are familiar with the Asch judgment experiments. Typically he presents three comparison lines and a standard line to six stooges and one non-stooge. The six stooges all say in succession that one of the comparison lines is the same length as the standard, when it is obvious that it is really longer. So this places the non-stooge in a conflict situation. He has just heard six other students say that these two things are objectively equal and it is perfectly plain that they are not equal. So what does he do? Well, under these pressure conditions, about a third of the subjects, e.g. college students, fold: They yield to social pressure and call out the wrong line.

Asch has wondered why some students yield and some do not. We found, quite surprisingly, that the non-yielders, the people who refused to yield under this pressure, are the ones with high achievement motivation. There is almost no overlap in n Achievement scores of the yielders and non-yielders.

A reason for this can be found in our research on the origins of achievement motivation. What kind of home background, what kind of childhood training is characteristic of the people with high and low achievement motivation? A very nice thesis has just been completed by Marian Winterbottom at the University of Michigan on this problem. It begins to explain how some of these things hang together. She was interested in the number of demands and restrictions that parents placed on their children, and at what age. She chose sons aged 8 to 10, and she interviewed their mothers and gave them questionnaire schedules to fill out.

What she found, to make a long story very short, is that the mothers of children with high achievement motivation made many more demands for independent decisions earlier than those with low achievement motivation. For example, consider an item she actually used, "Do you expect your child to learn his way around town by himself?" This is one aspect of independence training. All the mothers said they did require this of their sons. But the mothers of children with high achievement motivation said that they expected the child to know how to do this before the age of 8, which happened to be the median age at which the distribution of expected ages could be split. These mothers required more independence earlier; in other words, there was great pressure from these mothers for independent activity of various sorts—crossing the street by oneself, making friends, doing well in school, etc. All of these independence-training needs seemed to be required earlier by the mothers of sons with high achievement motivation. So I think you can begin to understand why the products of this kind of parental background would stand out against the pressure of the group in the Asch experiment, why they would be more at the democratic end of the Authoritarianism scale. And vice versa: you can see why the ones with low achievement motivation coming from a more protected background, would tend to be more dependent on other people of authority; why they would be willing to follow the crowd, even when it is wrong, and so forth. I need not elaborate.

Now, to turn to my last point, namely, the problem of predicting judgments of performance. This is a long way of saying "predicting grades," and I chose the long way on purpose. Predicting judgments of performance is no mean trick, as most of you know. I do not regard it as especially difficult in this case to predict performance, but to predict judgments of performance is quite a different matter; it is the criterion problem with which you are all familiar.

Actually, we have done a number of studies of the relationship between n Achievement score and grades in high schools, colleges of all sorts, and our correlations are sometimes high and sometimes low. I remember when we first ran this correlation, for a college sample; it came .51. We were so elated that we nearly sat down and sent a telegram to Professor Terman saying "Forget about your intelligence test; we can predict grades better with a 20-minute projective test." Well, it is a good thing we didn't, because we ran the correlation on another sample and the next time the correlation was zero. A healthy corrective for enthusiasm, the repeated experiment!

To summarize this research the way it stands now—the Educational Testing Service will straighten us out on some of these things, I hope—there is a median correlation of n Achievement with grades in the .20's with intelligence partialled out—significant, but nothing to get terribly excited about.

Let me mention what I think two of the main problems are in getting such predictions of grades. In the first place, how much does the criterion, namely, grades, depend on motivation in a particular case? We have found a case—and I am sure you know of such cases—where the correlation of Otis I.Q. with high school grades is 0.90. Can you expect any correlation of grades with motivation if this is so? You may find one, but you certainly aren't going to add anything to the prediction of grades that you get from the Otis I.Q. alone. Maybe the teacher just looked up the intelligence test scores and graded accordingly.

How much does the grade criterion depend on motivation? And how much on the teacher's idiosyncrasies? Langlie and others showed twenty-five years ago that grades in high school, at any rate, and I suppose in college, are correlated with teacher judgments of other personality characteristics, e.g., attractiveness, physical maturity, and other characteristics of that sort. So there is certainly impurity in the criterion—e.g., judgment of performance as compared with performance itself.

The other problem that has been very puzzling to me is whether or not it is really legitimate to parcel out intelligence. The normal way of proceeding is to correlate n Achievement with grades, intelligence with grades, and then parcel out the correlation of intelligence with n Achievement. There is always a positive correlation between achievement motivation and intelligence, and there ought to be, it seems to me. Take the extreme case. Whatever the native ability of a person, if he has no motivation to learn, he is not going to get a high intelligence test score. So it seems to me there ought to be some correlation between achievement motivation and intelligence test score.

There are two places where motivation enters into an intelligence test score: one in the accumulation of knowledge which he shows on the intelligence test or achievement test, and the other in the attention he gives at the time he takes the test. We know that people who have high achievement motivation will actually do better in the testing situation. So there is an intertwining here of achievement motivation and the intelligent measure. Is it fair then to partial out I.Q. in relating motivation to grades if we know motivation also determined the I.Q. to some extent? If we do, we are eliminating part of the effect that motivation has on performance. If we don't, we can be accused of simply finding a correlate of I.Q. which therefore ought to predict grades to some extent. It is a difficult problem to think through—the relation of motivation to performance and intelligence, but these are our contributions to it to date. At least we think we have a method of measuring motivation which should provide plenty of food for thought.

DISCUSSION

PARTICIPANTS

PHILIP ASH, EDWIN G. FLEMMING, CHARLES R. LANGMUIR, DAVID C. McCLELLAND, JOSEPH ZUBIN.

DR. ZUBIN: I believe we are all deeply indebted to Dr. McClelland for a very timely discussion of a problem that is facing research in personality. I have but three comments.

First, about the technique for eliciting achievement motivation which was used. We were not told exactly how it was done, but apparently it had something to do with the introduction of incentives for achievement in the one group and no such incentive for achievement in the other group. Of course that is a very school-like situation, and one wonders whether that kind of experimental eliciting of motivation bears a high degree of relationship to the wide variety of facets that motivation consist of. It may very well be an important aspect of motivation, but that it encompasses the entire variable that we regard as achievement motivation is doubtful. On the positive side, when one begins to tackle such a field, it is good to separate out the different facets, but whether the kind of achievement-motivation elicited in school is very important for achievement-motivation in life remains a very important question.

I also wonder whether the very simple test he used as an indication of motivation, namely the increase in rate of simple additions under incentive conditions, had previously been used by the Character Education Inquiry and by the Spearman School, for the measurement of motivation—I wonder whether that might not give as good a correlation with degree of achievement motivation present during the experiment as the dissection of the person's imaginative production on the TAT would yield.

This very simple task of simple addition may give you as much as the more complicated analysis.

As to my second point, the technique utilized by Dr. McClelland essentially consists of utilizing derivatives of the projective technic method. This is a very worthy derivative. We have been able to demonstrate not long ago in our own laboratory that when content analysis on TAT-like pictures are used tachistoscopically in the specific focused situation involving interpersonal relationships and the contents of the response for that particular variable are scaled, we find tremendous differences between the performance of individuals who are normal, those who are neurotic, those who are chronically ill mentally, and those who are only in the early stages of illness. The idea of using derivatives of the TAT or of other projective technics in a motive-focused manner and then scaling the results along the dimension under investigation is a very worthy one, and it has been applied not only to the TAT, but also to the Rorschach. Dr. McClelland's findings add much weight to this approach.

For example, when the content of the Rorschach is analyzed on specific scales for measuring dimensions of content involving such variables as cheerfulness, anxiety, sociability, etc., significant correlation is obtained, whereas as you know, ordinary clinical scoring of the Rorschach gives very low correlations with such personality variables. This whole method is part of a very healthy approach of trying to make sense out of the chaotic field of projective technics by singling out particular segments and focusing attention on the particular performance related to that segment.

The third point, and I think Dr. McClelland will agree with me, is that he has defined motivation in a very narrow setting. He has limited himself to what you might call, for lack of better terms, rivalry and competition, competition with norms, rivalry with others. But there is more to motivation than just that. Certainly professional motivation, if you limit it to these points of view, gives you only a very small part of the picture. What about cooperation as a motive? What about curiosity as a motive? What about altruism as a motive?

All of these motives are lost in the particular sector that Dr. McClelland has selected and I do not mean to say that therefore his work is not of value; it is of tremendous value. I believe however, that we should not be surprised at the low correlation between school-achievement and his achievement-motivation score because he has not measured motivation in all its aspects; he has taken two aspects of it which perhaps unfortunately the American scene stresses unduly. The other aspects of motivation may not be as strongly developed in the average person in our culture, but that they do form at least part of the achievement-motivation of many people cannot be doubted. Perhaps finding tests for measuring these latent motives may hasten their development.

Dr. McClelland: Let me make one comment. I am frequently accused of not doing things that I didn't intend to do in discussions of this particular method. Of course we didn't measure the curiosity motive, we didn't intend to. Also there are lots of other motives that certainly can be measured, using the same method I perhaps should have made that clear. I think the method that we used here of concentrating on one motive—and I would even agree that it is one aspect of one motive—is one that can be generally applied, and I think that is the main significance of what I said here today.

We intended to deal only with one aspect of motivation, and I think that the final test of whether that aspect was worth concentrating on or not is contained in the twenty or thirty relationships that we have between it and other important variables. That is the ultimate test of the usefulness of any analytic approach.

I remember when we first started doing this five years ago—six years ago now. Dr. Rapaport said substantially the same thing that Dr. Zubin said. He said the motive you arouse in the laboratory has nothing to do with achievement motivation in life; you are wasting your time. I am glad I didn't listen to him.

Dr. Flemming: I want to ask Dr. McClelland whether he has correlated this achievement test with practical achievement in the work situation, such, for instance, as the achievement of salesmen.

Chairman Bennett: I take it you all heard the question. Dr. McClelland says the answer is no.

Mr. Langmuir: Could Dr. McClelland give us any information about the variation in his measures of achievement motivation under different conditions of arousing it, or over an interval of time in successive tests of the same individual?

Dr. McClelland: This is a complicated question in a way. The stability of the achievement motivation measure as we now use it is not ordinarily high. People who are used to scoring intelligence tests are going to be alarmed at this. If you retest a person weekly, or six months later, you do not get as high test-retest or reliability cofficients as you ought to get, at least as we are used to thinking you ought to get. That is, they run probably as low as the sixties and seventies rather than up in the eighties and nineties.

There is another way of looking at the problem. It involves the whole question of the relationship between validity and reliability. The other way of looking at it is that if the measure was stable, or more stable, it probably wouldn't be as sensitive. In other words, intrinsically motivation is something which does vary probably more from day to day and week to week than intelligence does, since we more or less assume that intelligence is something which remains relatively stable over time; at least we try to measure it in a way which yields stability. Motivation, on the other hand, is something which I would say intrinsically varies more.

Dr. Ash: I have a question that relates to one asked previously. I gather that in its present form the test has not been used in the industrial situation, but I wonder, first, whether as it is used now, or as it can be used now, it might be appropriate in industrial testing. My second question would be, has it been used to observe relationships between need achievement and such variables as leadership behavior and group acceptability, for example, in line with the work done by Shartle at Ohio State.

Dr. McClelland: My answer is no, although there were some studies done that are a little bit relevant to this problem down at the University of Maryland. Field did a study on the effect of social rejection on the n Achievement score and he found very serious sex differences. I haven't mentioned the sex differences which appear with this test. They are very markedly dependent on the type of arousal, and this supports what Dr. Zubin said earlier. I agree with most of what he said. It is only that I was trying to do something different. I am afraid I sounded as if I did not agree with him. I do, because we know that different arousal conditions will produce different effects, and the big sex difference is a major case in point. For example, the achievement motivation score of women does not increase under our normal arousal conditions. We thought the women were very refractory; we tried and tried it, again and the men's score increased every time, by our scoring system, but the women's score didn't.

Field, at Maryland, did a study in which he rejected both men and women. That is, they were told that there was going to be a sort of popularity poll and that they were going to learn the results of it. He handed them back slips of paper on which it was clear that they were in the group or out of the group, rejected or accepted. Now, under these conditions, the women's achievement motivation score went way up if they were rejected. After we discovered this, we found that Else Frenkel-Brunswik had shown this years ago in her motivation study in which she found that high achievement motivation, as rated by teachers, was pretty closely correlated in girls with appearance, dressing, and things of this sort, with the social side of the achievement motivation, whereas in men it is more connected with leadership and intelligence (the factors referred to in our normal arousal conditions). Apparently it doesn't threaten a woman nearly as much to call her unintelligent as it does to call a man unintelligent.

I have greatly oversimplified the nature of the achievement motive. I want to say that. I am afraid my earlier remarks were not as serious as they should have been. There are all kinds of achievement motives. We know, for example, that there are some people who are characterized primarily by a hope of success, others by a fear of failure. I did not have time to discuss all these variations.

There are some whose achievement motivation is focussed on athletics, or playing bridge, or being a Don Juan. So I certainly have to agree with Dr. Zubin that the motive is much more complicated than what our simple, over all index shows. The research problems remaining are very great indeed.

LUNCHEON SYMPOSIUM

Trends in Public Opinion Polling Since 1948 and Their Probable Effect on 1952 Election Predictions

Trends in Public Opinion Polling Since 1948 and Their Probable Effect on 1952 Election Predictions

FREDERICK F. STEPHAN

WHAT ABOUT THE SAMPLING? A BIT OF PSEUDOHISTORY

I SUPPOSE all of you are well aware of the fact that the sampling methods that our polling organizations are using this year have evolved from quite primitive beginnings. First of all, of course, there was a plant life period or stage in which a newspaper reporter asked the people he found around him how they were going to vote. He sent out his roots like a plant and got what he could from the spot where he happened to be. Then came the dinosaur stage, the huge mail canvasses of ten million or more post-card ballots that were sent out by the "Literary Digest" and similar surveys made by other publications, ending up in the swamps. They were bogged down by the mere size, and if anything they proved that size alone is not sufficient for survival.

Next, if we can jump a million years, more or less, came the dawn of civilization, and with it came men who had at least the rudimentary beginnings of an alphabet. They roamed over their hunting grounds, capturing big, lumbering elephants and stocky, stubborn donkeys in order to put on a great race. They laid heavy bets on the spectacle and gave odds, and everyone tried to dope out the outcome beforehand so that he could bet on a sure thing.

They took their primitive alphabet and marked the fattest animals A's, the good and plump ones B's, the middleweights C's, and the scrawny ones D's, so they would have a fair mixture of economic levels and have a good race. In like manner they gathered a proper mixture of males and females, old and young, cave dwellers and denizens of the forests. They did this every four years and they had great fun. But in the end it availed them naught, for after they flourished for a while, they grew too bold, and in 1948 B. C., they plunged with the biggest bet of all time. And when they lost their shirts, and all their following had likewise, there arose a great wailing and a loud cry, "Oh, what have we done, or failed to do, oh, Lady Luck, that you should desert us now?"

And they approached the soothsayers, as did all the multitude, and the soothsayers said, "You were not careful enough in most of what you did, and you should have done much that you neglected to do, but most of all, you didn't pay attention to Mother History, or learn your lessons about the probable error, and the last minute shift, and the mysterious ways of the undecided and evasive, and the wiles of the editor who wants you to do your stunt way out on the limb, and the heartbreak of the photo-finish." And they went away sad and repentent.

Well, now, the year has come around for the next race. The pollsters have managed to gather the animals again and all the followers of Mother History are asking, "Will they go off chasing Lady Luck again? Have they learned their lesson? What about those pitfalls? What about the sampling?"

Well, now fellow cave dwellers, no one can tell whether the pollsters will tag after Lady Luck until tomorrow, or perhaps Monday morning when the final call comes for placing the big bets. They say they won't be betting this time, maybe never. Here (holding up a letter) one of them says. "Us? We are not predicting."

They hope that other people won't use their reports as a racing form and lose money on foolish bets, but still they say, "We will tell you all we know, everything we know about the animals, and then it is up to you."

They have gone farther into the dawn of civilization and they have learned a great deal about numbers and counting, at least on one hand (counting fingers), 1, 2, 3, 4, 5, and they have learned a new religion called "Probability Sampling." Some of them have embraced the new religion while keeping a few of their old pagan beliefs. Some others are trying to use the same old prayers and magic again for they hear that the new religion is a strict master and that it exacts a heavy price before it will help them in any way. Whatever their present faith and doubts, none of them seems confident of his dope on the race, or hopeful about the benign intervention of the supernatural. They are saying, and will probably continue to say, "it's anybody's race."

That is as far as I will go in predicting what Sunday morning's or Monday morning's final releases will be like.

Well, what about sampling, then? Some of the state and local polling organizations, and quite a number of newspapers, appear to be using very much the same methods, at least so far as sampling is concerned, that they used in 1948.

There are, of course, the same old half-serious stunts: the chicken-feed poll is on again, but we don't know whether it is the farmers or the hens that are making the big decisions about who is going to be elected. The cigarette poll is on. It samples smokers, ignores non-smokers; undoubtedly a bias right there. The taxi driver poll is on, the barber poll, and no doubt the astrologers are busy, too.

The important problems of sampling do not center in these side-shows in the election circus. Neither are they to be found in those really serious canvasses that are operating in certain instances as if nothing happened in 1948, and in other instances as if the important thing is not to try to pick the winner, but to really find out in a genuinely scientific way something more than we presently know about political behavior and the way in which people make up their minds.

They are to be found in the more noteworthy surveys of opinion and election behavior that are now being made, some of which have not come to our attention because their results will be reported more deliberately after the election is over. However, some of the better-known polling organizations are also contributing more than they have previously to the more serious, long-range study of election behavior. Therefore, even though there may be few forecasts, the sampling problems associated with their work is still very important, for we must appraise the results of the polls in so far as they offer any possibility of increasing our understanding of how people think on issues and how they decide to cast their ballots on election day.

The analysis of the details of the sampling operations as they are actually carried out by the pollsters is a major undertaking that none of us has attempted so far. I would like to stress that. It is not something that you can do in a day or that you can do without going through a great deal of material that is available only in their offices. Some of the material we would need isn't even available there. To judge how the sampling operations are really working out now compared with previous operations was difficult enough in 1948; it is more difficult today, I think. Nevertheless, I will attempt to make a few general observations on the methods that are being used by the more prominent of the polling organizations. They will gloss over the details and give us just broad outlines.

The principal changes that have been made since 1948 in the selection of the sample of people to be interviewed have been made in the direction of assigning to interviewers a selection of city blocks and specific directions on how to do the interviewing in each block. The instructions specify the starting point on each block and tell how the interviewer should count off a designated number of households from each selected household to pick the next household in which to seek an interview. There are instructions about how an individual is to be selected within each household that is thus chosen for the sample.

There are also some general arrangements designed to get more of the interviewing into the period when people are home from work in order to reduce the losses that occur in daytime interviewing. This procedure replaces the old quota sampling procedure in which the interviewer had a relatively free choice of respondents so long as he satisfied certain quotas assigned by economic level and sex and followed certain general instructions about obtaining a representative group of respondents to interview. The relatively new procedure of "block sampling" is actually forcing interviewers to go into areas in cities that they had avoided before, or missed altogether. It will probably remove much of the bias in economic level and education that was characteristic of previous polls. How much, we can't say, but it seems to be a direct consequence that it should have that effect.

However, there is a general disposition among polling organizations not to require interviewers to make additional calls when they find no one at home at the first attempt. In some instances there is a provision for substituting a neighbor, but in other instances no attempt is made to replace or to regain interviewing attempts that are unsuccessful in the first instance. In addition, the older quota sampling methods are still employed in some of the rural areas or in other situations in which the block sampling procedure is difficult to apply.

There are some other types of sampling that are being used, such as Gallup's pin-point method, but they represent supplements to the main samples rather than the principal sampling procedure itself. If there were time enough and you were interested in such details, it would be quite appropriate, I think, to examine the ingenious attempts that are being made to find a way of sampling that is not as costly and troublesome as people think that probability sampling is, and yet that avoids some of the weaknesses of the older methods.

Back in 1948 a special committee of the Social Science Research Council made a comprehensive review of the polls shortly after the election. It concluded in its consideration of the sampling problem that the available evidence was not adequate to measure the extent to which quota sampling contributed to the systematic errors in the 1948 election predictions. On the other hand, the examination of the sample educational distributions suggested that there was a considerable systematic error in most of the quota samples, although the amount of it, the magnitude of it, could not be measured.

The report also held that there was a possibility of improving the sampling methods, but emphasized the fact that numerous factors other than the sampling error contributed to the gross error of predicting the division of the vote among the candidates. It warned specifically that the use of probability samples will not in any way guarantee that one can predict elections. Hence, we may expect that these current changes in the direction of probability sampling will improve the accuracy of the polls somewhat, but will not enable them to succeed where they failed in 1948.

In the absence of a definite analysis of their accuracy, we must assume that the current percentages are still subject to a degree of error, from sampling alone, of the order represented by a standard deviation of perhaps two or three percentage points. This is little more than a guess based upon the past performance of the polls and what we know about the general outlines of the sampling methods now.

Including other sources of error the gross or total error may be of the order represented by a standard deviation of 5 to 6 percentage points. Now, this doesn't mean that they can't be, as we used to say, "right on the nose," but as you all know from the applications of the theory of error in testing and related fields, what is important is not the possibility of being exactly right and having a zero error; it is what the long-run experience of a variety of errors leads us to expect. The guidance we can get by assuming some approximate value of the standard deviation is an important element in reaching a sound judgment about the meaning of the polls. These guesses may well be exceeded in the case of samples or sample results that are based on fewer than, say, a thousand respondents or polls that are subject to more than the average degree of error.

Therefore, when you examine any of the results of the polls, please increase each percentage by 5 percentage points or more, and also subtract from each percentage the same number of points, then look at the two figures you get and draw your conclusions from the assumption that if you continue operating in this way the percentage you wish you knew will be caught between these two figures about twice as often as it falls outside.

You know how to adjust this if you want to change the odds of trapping the supposed true figure you are seeking. And then don't forget that about one-third of the time even this crude way of making allowance for inaccuracy will lead us to underestimate the actual gross error in the particular percentage we have before us.

What about sampling, then? It has been improved, but not as much as it could be. It exhibits all the practical problems we encounter when we attempt a large house-to-house survey in any field, but this isn't the main reason why the pollsters cannot tell you clearly who will be elected, or very accurately how different groups of the population differ in their reactions to issues and candidates. That is a story for the next two cave men to tell you.

Trends in Public Opinion Polling Since 1948 and Their Probable Effect on 1952 Election Predictions

HERBERT HYMAN

INTERVIEWING

AFTER THE polls failed in 1948, my colleagues here, and I, took part in an investigation of the polling organizations. This time it appears that we are taking no chances. We are investigating them even before they have had any time to fail. I might say that the three major polling organizations have been very courteous to us and allowed us, in a sense, to do this detective job again and we are most appreciative.

During that earlier investigation, the story went around of the interviewer who wrote to her agency and remarked that she knew why the polls had failed. She described her experience in interviewing on the elections and said that she had this strange experience: she kept running into respondents who continually reported that they were planning to vote for Truman. This happened so often that she knew something was wrong. She surmised something was wrong with her interviewing or her sampling, and so she threw out some of those cases and did some more interviewing until she found enough Dewey supporters.

I suppose that this story—which, no doubt, was invented by some wit rather than being the real truth—is a kind of dramatic illustration of the contribution that the interviewer conceivably might make to the success or failure of the polls to predict an election. Obviously the sampling and the research design and the analysis can be expert, but in so far as the raw data that are collected are inadequate, of course, the error is implicit in all the later predictions.

This possibility of interviewer error is perhaps the reason why I was assigned the topic of changes in interviewing methodology since the '48 polls. In actuality, there is very little to be said in the way of anything new on the problem, for the polls have made no real, radical changes in their interviewing procedures or their interviewing staffs and the detailed story of this aspect of survey research can still be found in the *Social Science Research Council Bulletin No. 60* on the 1948 election investigation. I might say that in the *Bulletin,* there is a detailed chapter which attempted to estimate the magnitude of error created through the interviewing process, and, as Mr. Stephan pointed out with respect to sampling, it is impossible to determine exactly how much the interviewer component of error was responsible. There was some putative evidence that it could account for, let's say, a rather small part of it.

Now, this absence of change in the '52 research procedures might be regarded by you as negligence on the part of the polling agencies. While I do not want to condone this pattern on their part, I might describe to you certain features of the interviewing procedure in one of these major agencies which help account for the fact that there has been a persistence of the same methods. Incidentally, this information might be of some intrinsic interest to you apart from its relevance in explaining the situation.

Interviewing in these survey agencies must not be conceived in the image of interviewing in an academic research study or in the image of interviewing for clinical or psychometric purposes.

Interviewing in the survey agency represents a very massive field operation, conducted generally by part-time, rather poorly paid employees. Their rates of pay run between a dollar and a dollar and a half an hour. The number of field workers engaged in an election survey would vary with the agency, but would run into perhaps two hundred or so interviewers. Certain organizational features of the research agency itself hinder any change in the composition, geographical distribution, or operation of such a staff, no matter how well advised such changes may be for a given election prediction survey.

First among these is the fact that all the election polling agencies conduct these particular polls as a side line. Their major work, except for an occasional trauma every four years, consists of market research and a variety of opinion studies. The character of the field staff operation is and must be essentially dictated by these more continuing needs rather than by the specific needs associated with effective election predictions. Thus, for example, the sample designed for predicting an election might call for interviewers in certain areas in certain strengths, as, for example, in making an estimate of critical states, which strengths would otherwise not be needed for the rest of the year or for another four-year period.

Similarly, the election situation might call for a certain political composition in the staff, because of the possible ideological bias of the interviewer, but this same political composition may be irrelevant for most other market research purposes.

Radical alterations in the composition of this staff due to replacement or firing or necessary additions involve a rather considerable expenditure. While the cost of recruiting and training and supervision of a single interviewer is difficult to determine, we might, for our purposes here, set the figure at $75 per unit interviewer, which would be a ridiculously conservative estimate. Now, this figure may appear negligible to you, but, when you multiply this two hundred times, you find that the usual agency has an equity of perhaps $15,000 in its current field operation, a property not easily jeopardized, considering the fact that these agencies are commercially run. But even where change may be called for and organizational factors such as I have mentioned are ignored, there is a great difficulty in making fundamental changes in the composition of survey interviewers. For example, it is interesting to note that on the present continuing permanent field staff of the Roper agency, there is only one lonely male interviewer out of perhaps 250. All the rest are women.

This situation, I can assure you, does not represent the agency's libido at work. This is a product of larger institutional factors that affect the type of individual that is available in the labor market from which interviewers can be recruited for survey research.

In the course of a detailed investigation of interviewing and survey research that the National Opinion Research Center has been engaged in under SSRC and Rockefeller Foundation auspices, my colleague, Paul Sheatsley, conducted a very intensive study of this labor market for interviewers. He notes such facts as the following: while there is some variability within this market, and the staffs of different agencies show different profiles, there is a modal type of interviewer available for hire. No matter what agency or which time period is studied, the college educated comprise about three-fourths of all the staffs, women at least two-thirds or more.

The Negro interviewers, whom NORC was able to hire over the past twelve years, are an educational elite. They are far better educated even than our white interviewers, about one-third of them having done post graduate work beyond college. This is quite interesting when you consider the fact that they are interviewing by and large a segment of the population with far less education than they have.

One other finding by Sheatsley is of interest. The rigidity of this labor market was examined by comparing the characteristics of those interviewers hired before World War II, during the war, and in periods since the war. Despite the massive population shifts due to wartime factors, the characteristics of the field staffs hired were fairly stable over all this time. The same stability is demonstrated, no matter which supervisor tries to recruit interviewers in any area. For these organizational and institutional reasons, the composition of the interviewing staffs used show no major change. Nor has there been any major change in the training of these interviewers or in the actual conduct of the interview.

Such procedures of training are fairly standard, fairly rigidly institutionalized in the agency, are regarded by them as working moderately well to insure quality, and they reason that the problems of prediction relate much more to the realm of sample design or to the realm of conceptualization and analysis of voting preferences. These are the problems my colleagues are addressing themselves to.

On the score of training and supervision, I should report, however, improved methods of quality control. For example, Crossley has developed a procedure which he started in 1948 of checking the performance of each interviewer by comparing each interviewer's results on a series of demographic characteristics with criterion data on that characteristic for the same sample point. Any major discrepancies between the results for that interviewer and the criterion data imply either error or, what is worse, cheating, that is, that the interviewer fills out the answers himself in the privacy of his home. Under such conditions, Crossley institutes some disciplinary action against this interviewer.

Roper similarly has expanded a system of controls involving regional supervisors who report on the quality of performance of the interviewers under them through direct observation and through filling out detailed rating sheets. A description of that procedure is given in detail in the summer 1952 issue of the *Public Opinion Quarterly*. Consequently, there is reason to believe that while the interviewers have not changed in character, they are under somewhat better control.

There is also reason to feel that they are a pretty highly experienced staff for the type of field problem they are encountering. Roper, for example, had one interviewer who recently died after a length of service of sixteen years. Crossley has at least fifteen interviewers as of the present who have had lengths of service running between 20 and 27 years. These people are obviously of considerable experience.

With respect to the actual interviewing procedure that is used, it should again be noted that interviewing in the survey must not be regarded in the same way as other types of interviewing procedure. The procedure that is used in election predictions is essentially predetermined by the standardized questionnaire developed and by the explicit accompanying instructions rather than by the discretion of the interviewer.

The interviewer is, in a sense, much more a machine rather than a professional person given freedom to exercise his judgment, that is, apart from the choice of respondents in a sampling design.

Changes in interviewing procedure for election purposes are really much more the province of research design. On this score we might note a few changes in design that in turn affect the interviewing assignment. For example, in 1948 there was reason to believe that last minute shifts were of considerable significance and would have to be treated in future research, and so there is greater emphasis this time on telegraphic surveys which involve the interviewer operating under conditions of stringent deadlines, fast interviewing, and a return of the results by telegram. In a sense, this improves the design but creates certain additional possibilities for errors due to hastiness and pressure.

Similarly, Crossley in '48 initiated some research into filtering out ineligible voters and uninterested voters. This procedure of filters seemed a very good procedure, and he reports that he has developed it further. Again this places upon the interviewer additional difficulties, because these filters must be treated differently in different parts of the country. Eligibility requirements for voting vary in the most capricious way from place to place, and the interviewer in exercising these filters in the interview must evaluate them differently from area to area.

The other major improvement in question design which affects the interviewers, notably in Roper's work, is the use of batteries of issue questions which attempt to define the constellation of attitudes surrounding the preference which either make that preference sturdy or make it precarious because of conflict between desire to vote a candidate in and desire to see certain ends achieved with respect to issues. This naturally creates more difficulty for the interviewer particularly because while Roper can easily, in this way, see the constellation of attitudes surrounding the preference, he must also be in a position to evaluate the hierarchical importance of different issues within this constellation, which involves basically open-ended interviewing.

These are some of the question changes that in turn must be implemented in the interview situation. Of course, apart from the specific procedures of interviewing, there is the general problem of rapport and inter-personal relations with the respondent. On this problem, you might think that the memory of the 1948 fiasco would be carried by the population and impede effective interviewing. However, trend data collected by the NORC since 1947 indicate that the permanent decrement in public confidence is negligible. In November '48 there was an all-time low in such public confidence, but a figure collected this month indicates that the polls have regained their status with the public; only about two percent of the national sample actually feeling really hostile to the poll, and about six out of every ten reporting a favorable view.

More than this, some of the general psychological difficulties associated with the interview situation in '48 seem to be less operative. It was quite common in '48 to obtain reports from interviewers of a hidden Wallace vote which was not declared out of fear of stigmatization. There was even an occasional report then of annoying confusion between the names Truman and Thurmond.

This general problem of evasion as in the case of the Wallace vote and consequent response error does not seem so present now. Only sporadic difficulties are being reported. The interviewers are remarking on the high level of respondent interest, the willingness to talk, and even the fact that women, normally very apathetic in political polls, are alert and interested.

This would seem to be a brief account of the changes, or rather lack of change in the interviewing aspect of the election polls since '48. I am not implying that the polls have made no changes elsewhere, or that they should make none; they have serious problems elsewhere, and perhaps some minor ones in the interviewing field. But the problems that are most crucial seem to lie elsewhere in the research process and the agencies show sound judgment in allocating more of their energies in those directions.

Trends in Public Opinion Polling Since 1948 and Their Probable Effect on 1952 Election Predictions

SAMUEL STOUFFER

ANALYSIS

Dr. Samuel Stouffer: I think that everybody in this audience has a serious professional stake in what the polls do in this election. Let's not forget what some of our friends and critics who have no use for psychometrics or for quantitative methods in general had to say in 1948, and how some of them tended to draw the conclusion that human nature and human behavior is intrinsically unpredictable. Hence, it could be inferred, we cannot even predict on an actuarial basis whether people will do well in college on the basis of previous tests.

This type of attitude was one which was lusciously enjoyed by some of our colleagues in the humanities, and such distinguished scientific journals as *The New Yorker*, the week after the 1948 election, came out with choice statements expressing gratitude to the pollsters for clouding up the crystal ball and expressing respect for the American public for telling one thing to the poll-taker and doing the opposite in the voting booth.

I want to make a few brief points on the very large subject of how the polls are handling the analysis of the data they are collecting. First of all, I want to say that I think the integrity of the major pollsters is beyond question. I think they deserve a great deal of credit for courage for behaving cautiously, because it would probably be better from the standpoint of public reaction if they said they were sure that Eisenhower was going to win or that Stevenson was going to win than to hedge. But they intend to speak definitely only if they are convinced that their data point that way; and their data are not likely to point conclusively enough one way or another to make it possible for them to make a definitive statement.

Why is this likely to be the case, apart from the problems of sampling and apart from the problems of interviewing?

TESTING PROBLEMS

The first point I want to make has to do with the peculiarity of our electoral system. In 1948, even if the polls had been right on the button, with an error of practically nothing, a forecast would have been a gamble. If Dewey had got about half a per cent more of the popular vote, he would have won the election. No poll is going to be that close. If Truman had received one per cent more than he got, Truman would have won by an electoral landslide of four to one, which would have been one of the most unprecedented electoral landslides in American history. That was because the vote was so close in all of the key states.

Therefore, it is quite clear that even if a poll is extremely accurate, our electoral situation may make it impossible to predict an electoral vote with any precision at all. Today the pollsters are telling the people that in every way they can possibly do it. But they didn't tell the people that enough in 1948. Some of them said it, but they just didn't say it strongly enough. Now they are saying it, and of course others are laughing and saying, "Well, the pollsters are just not going to take any chances." Actually, of course, it was relatively easier in the Roosevelt period. There is a good reason why it was easier, and that was shown by the fact that the polls in the Roosevelt elections showed a relatively small number of undecided voters. People pretty well knew what they thought about Roosevelt. He was either the great hero who had saved the country and later the world, or he was "that man," and the number of people who probably made up their minds during the campaign was relatively small. There was not much evidence of fluctuations after the first of September.

The 1948 election represented something very different indeed. There was a large number of undecided voters—twice as large, the polls themselves show, away back in September, 1948, as compared with earlier elections, but experts did not take it too seriously. Previous studies of the undecided voters showed that they tended to go about like the rest of voters did; hence the tendency was to neglect the undecided voter.

The other thing that was neglected was the possibility of some last-minute changes in the attitudes. The 1952 election has followed a course up to, I would say, a week or ten days ago, according to the polls, that is very similar to the 1948 election. You have a big start, apparently, for Eisenhower gradually being dwindled away, and then you come to today— a few days before the election. Now, at this point, what is going to happen? This has the pollsters tearing their hair and they are making very careful polls this week, trying to see whether or not there is any evidence of a very sharp pro-Democratic trend which happened in the last week or whether the difference in the campaign procedures this year, particularly the Republicans' efforts to maintain their momentum to introduce the Korean issue with all the power they know how, will prevent any trend towards Stevenson from continuing.

One cannot make empirical generalizations with confidence that what happened in one election necessarily will be repeated in another.

FIGURE I

	Probability of Voting				
	Almost certain not to vote	Quite likely not to vote	50–50	Quite likely to vote	Almost certain to vote
Definite Eisenhower					
Leaning Eisenhower					
50–50					
Leaning Stevenson					
Definite Stevenson					

But if the trend for Stevenson continues, the pollsters realize that they could be facing on election morning a figure which shows the two candidates about fifty-fifty in the popular vote. If, however, the trend does not continue, the election will probably show Eisenhower with a margin of popular vote, which still doesn't mean he necessarily would win in the electoral vote.

This year the pollsters have sought explicitly to take into account two different kinds of uncertainty which enter into responses. These are illustrated as variables in Figure I. On the horizontal axis we have varying degrees of probability of voting. On the vertical axis we have varying degrees of enthusiasm for the candidates.

Now the polls can fill in each cell with frequencies. If you cut the horizontal axis somewhere near the middle, on the assumption that only the more probable half of the voters will vote, you can consolidate all the frequencies to the right of the cutting point and come up with a figure as to how the probable voters are leaning. But in the middle of the vertical dimension are a block of voters who haven't made up their minds. We can ignore them and base our estimate on the two upper and two lower tiers or we can make some assumptions about them. In 1948 people in the middle blocks voted for Truman—but it is not safe to assume they will vote Democratic this year, even though their characteristics are Democratic.

And there are additional complications. Many of those leaning toward Eisenhower are normal Democratic voters, some of whom say they are for Eisenhower but prefer the Democratic Party. Will they vote for Ike? Such people in 1948 who said they were going to vote for Dewey tended to swing back into the Democratic camp in the last week or two. *This time* they may mean what they say. But we cannot be positively sure and that is why the present caution of the pollsters is eminently justified.

Finally, I want to say a word or two about one of the most important things that the pollsters are doing, and that has to do with the analysis of the cross pressures which are present in this election. I am sure that you and I agree that prediction, particularly predicting a national election, is a pretty dangerous thing and it can have a boomerang effect. On the other hand, the polling data with all its intrinsic errors represent the very best information which we have about the trends in public opinion and about the ways in which issues impinge on various classes of our population.

I think we can be very confident from what the polls have told us that the majority of voters in this election—we can have five or ten per cent error and still be all right on this—really think they would personally be better off if the Democrats won the election. I think we can also trust what the pollsters tell us when they say the majority of voters think the Republicans can handle such problems as Communism and corruption better than the Democrats. You have people in basic conflict who think that personally they would be better off if the Democrats won, and they do not like the Communism and corruption business. The Irish Catholics are a very good example. And so the polls provide data which make it possible to take Irish Catholics who say they are going to vote for Stevenson, Irish Catholics who say they are going to vote for Eisenhower, and examine by correlational procedures their responses to a variety of questions, including some open-ended questions on what they think are the most important issues. The variety of questions asked will give us a better picture than any other procedure known as to how those issues impinge on various segments of such a population. It won't prove anything in terms of causation, but I think we can make inferences from it that are safer than any other kinds of inferences would be.

I do not want, however, to sell the prediction element short. In spite of all the difficulties involved in this matter, I think one can say something with a good deal of confidence about the directions in which the vote is leaning in certain states, and that may be useful too. I like to look upon this kind of prediction as a little like the job of the Weather Bureau in its longer range forecasting, where it is forecasting the weather for the week-end at the beginning of the week. Now, the Weather Bureau is going to make mistakes; it will make serious mistakes. It is making its predictions in terms of probability. It has a job trying to educate the public to realize that these answers aren't definite, but that they represent a probability statement which is genuinely better than the guesses made by somebody who sniffs with his nose and feels out the weather. The only trouble is that the public has been mis-educated to some extent and expects the Weather Bureau to say that there is going to be exactly 2.4 inches of rain or 1.2 inches of rain. That kind of prediction the pollsters can't make. But I have got a good deal of confidence that polling procedures are going to become more and more acceptable and that there will be public support for the improvement of the procedures, for I have confidence in the integrity of the pollsters.

PANEL II

Techniques for the Development of Unbiased Tests

Techniques for the Development of Unbiased Tests

IRVING LORGE

DIFFERENCE OR BIAS IN TESTS OF INTELLIGENCE

FROM TIME to time, scientists need to reappraise the concepts of their science, their methods of measurement, and the application of their knowledges for the general good. Psychologists, during the nature-nurture controversy, have had to reevaluate not only the concept of intelligence but also that of environment. For more than fifty years, they have been revising the *meaning* of intelligence, the various tests and procedures for its estimation, and, more especially, the implications of the evidence from tests for the understanding of children and their achievements. And, of course, they have critically reviewed the applicability of general, and special intelligence, tests for the selection, classification and guidance of individuals.

Psychologists, as well as educators in the fulness of time may feel obligated to the authors for "Intelligence and Cultural Differences." For again, they have asked them to reconsider the meaning of test intelligence. As contemplated, the book has motivated anew serious reexamination of intelligence and of intelligence-tests. Perhaps the authors, too, intended that some psychologists should become emotionally disturbed by the use of "differences" in the title in contrast with the use of "bias" within the text. Such feelings of disturbance may arise when such psychologists think of *bias* as some procedure by which some person "with malice aforethought" consciously prejudices a method of measurement to support an unfavorable (or favorable) opinion about persons, things or ideas. Few objective psychologists report "differences" for the purpose of proving a bias or a disparity. Most studies of individual or of trait differences, beginning with Galton and including Eells, have provided the evidence that measurable differences between groups exist. In test-intelligence, in particular, whether general or specific, *differences* have been found between groups classified by sex, and by age, and by education, and by geographic origin, and by occupation of father, and by cultural background, and by socio-economic status. Indeed, differences have been found in test-intelligence between groups classified by body-type, and by physical health, and by personality structure, and by nutritional status, and by family unity. Such reported differences from tests of intelligence have made test-makers as well as test-users increasingly aware of the multiplicity and intricasies of factors related to test performances of individuals and of groups. Not only are differences affiliated with groups, but they are affected by environment. Inadequate stimulation, within deprivational environments, may affect performance negatively. Indeed, we do now recognize the interactions of heredity as endowment and environment as opportunity for each maturing individual. Children, who during their early years, are deprived of linguistic, and of social, stimulation, as a group, do poorly in test-intelligence, and indeed, often are inadequate to cope with the range of adjustments the environment demands. The fact of "differences" is well-established: test performance reflects the specifics of environmental opportunities of training, of experience, and of stored achievement.

Test-users, have been instructed, over and over again, that an individual's test score must be interpreted always in light of an understanding of the variety of factors and conditions that are related to measures of intellect. Psychologists have provided normative data for a variety of groups because they know differences in test performance are related to sex, age, grade-placement, and socio-economic status. Furthermore, they have cautioned that a child's motivations and physical well-being do influence test performances.

Inevitably some users of tests neglected to profit from the tutelage. They wilfully treated test scores as absolute determinations about individuals, or, even, groups. Others, of course, failed to appreciate fully the range and interaction of circumstances that affect test-performance. To overcome such perversity and such ignorance, some psychometricians tried to be quit of the *bias* of the test-user by attempting to eliminate the *differences* from the tests.

Usually, the attempt to make an *unbiased* test of intelligence is an attempt to reduce some kind of group *difference* to zero. For instance, it is well-known that boys and girls (and men and women) perform differently on tests of verbal, and of numerical, content and process. For fear that the *biased opinion* that women are superior to men should predominate, psychometricians, for upwards of a half century, have reduced "differences" by addition. All of us are fully aware that to overcome the obtained verbal superiority of women, the test-maker adds a sufficiency of numerical reasoning items to make the average total score of men equal that of women. *No difference, ergo, no bias.* Fortunately, there still are differences between the sexes.

Partial justification, indeed, does exist for such a procedure. In general, a test-score that is based on a composite of many kinds of intellectual processes and contents does give valid (and reliable) estimates about most person's potentialities for success with the kinds of ideas and skills taught in schools. The emphasis should be on "most": many, however, may be misappraised because a score from many different tasks will fail to reveal the facts about differences within the individual's mental organization, and hence, by extension, fail to give information about differences in the mental organization of different groups. Of course, to apply Galton's suggestion of appraising many "shafts" does require more time than most test-users are willing to expend. For practical purposes, then, psychometricians have accepted either Binet's theory about the unitary character of intelligence, or Spearman's demonstration of the pervasiveness of "g." The consequent acceptance of the single index of mental-age, or intelligence quotient, or an intelligence score led to expectations that these scores were the absolutes about a person. They are not. The results of factor analysis have proved the need for the measurement of different aspects of intelligent functioning. Basically, differential aptitude tests attempt to measure "differences" as differences.

The measurement of "differences," however, is both costly in test construction and expensive in testing time, so that the *single-index* score will exist for some time to come. It must be recognized that most, if not all, so-called unbiased tests of intelligence are such *single-index* appraisals.

Another method for attempting to produce an unbiased test score is to try to reduce group *difference* by subtraction. Essentially, instead of adding items to conceal a difference, this method removes the items that produce the difference. The research reported in "Intelligence and Cultural Differences" deals with a technique for discovering the kinds of items in some current tests of intelligence that differentiate between some socio-economic groups. Eells, as a matter of fact, has found a significant relationship between measures of social and economic status and measures of intelligence. He realized, as careful workers had before him, that, on the average, the test-intelligence of groups with lower socio-economic status scores was lower than that of those whose status was higher. The fact of socio-economic differences in test-intelligence is reconfirmed. The implications of those facts, too, had led to the development of social inventions to reduce the environmental differentials which may affect the test performance of the socially and economically less privileged. Indeed, the full history of American educational legislation and practice from the "Old Deluder" Act to the contemporary requirement of compulsory schooling for all, illustrates the dynamics of democratic social engineering. Eells and his co-workers, however, took a different view of the facts.

They, apparently, assumed that the individuals in the various socio-economic stratifications were *equal* in intelligence. Hence, if any differences were found, it must be the test or some kinds of test-items that produces the differences. Thus was created the logical dilemma: *difference, ergo, bias.* In avoiding the one horn, psychologists must inevitably be embarrassed on the other. In facing the alternatives, however, educators and psychologists must be aware of general nature of the procedure. Eells, having established that differences obtain in test-intelligence between groups that they assigned to socio-economic strata, proceeded to select two samples at either extreme of the status score range, namely, children of old American stock who were either classified as of very high or very low status by the credits on the "Composite Index of Status Characteristics." He, then, made an item-analysis of a large portion of the tasks in the several intelligence tests that the individuals in each extreme had taken. Since the median percent of correct responses for the High Status group was about 81, and that for the Low Status group was about 70, it must follow that a plurality of the items will favor the High Status groups. This, indeed, was established by the analysis. The interesting finding, however, is that the differences in item performance between the two extreme status groups has "a direct relation to the form of symbolism in which the item is expressed." The High Status group is favored most on verbal items, but the gradient of difference becomes less and less for items "based on meaningless number combinations" and approaches zero for items involving pictures, geometric-design, and stylized drawing. Apparently, the discovery of such a symbolism difference *suggests* that a culture-fair test could be made of those tasks that minimize verbal processes and that favor those that require the manipulation of numbers, geometric designs and pictures. Such a test of such tasks, of course, can be made. But, if it were, what would it measure? It seems excessively trustful to put reliance only on such items that fail to distinguish between demonstrably different status groups. Some criterion about intellectual functioning, other than the one that the items make for no diversity, seems, at least, a psychological prerequisite. Certainly, within each extreme, variation in test performance must have been symptomatic of intelligent behavior that, to a very large degree, was a consequent of differences in ability or aptitude.

If such an unbiased test were produced by subtraction, it neither would be a test of intelligence nor would it give any evidence about the impact of status or culture on test performance. Certainly, Eells and his co-authors had methods available for item selection that would have maintained some relation to a criterion for intelligent functioning while minimizing the impact of status or culture. At least, partial correlation would have led to the making of a culture-fair test without losing the appraisal of intelligent behavior. At best, Eells' method could produce a test—but it would be a matter of conjecture as to what such a test measures. Clearly, the evidence from the many so-called non-verbal and non-language tests suggests that what they measure is different from what is measured by the so-called verbal tests.

Of course, the administration of the same verbal test to groups maturing under different language experiences would favor the group for whom the test language was their own vernacular. Test scores from a verbal intelligence test designed for Chinese would certainly put some Americans at a disadvantage. Indeed, not only will groups perform differently if they are separated widely by their languages but also if they have developed different cultural attitudes and values. Many psychometricians have endeavored to produce tests which are culture-free. From the days of Army Beta, attempts to remove the differences attributable to culture have been ingenious although not fully successful. To the long line of such tests, including Dodd's International Group Mental Test, Cattell's Culture-Free Intelligence Test, Spearman's Visual Perception Test, and the Multi-Mental Non-Language Test, should be added Rulon's Semantic Test of Intelligence. Each one of these ventures to achieve an unbiased test by substitution. Since the fact of different cultural and linguistic background prohibits the use of the language of any one group or a language common to all groups, the test-maker attempts to appraise intellectual performance by the manipulation of objects, or of pictures, or of designs or of numbers. The tasks, set by the psychologist, require intelligent behaviors of perception, selection, generalization, and organization. In cross-cultural comparisons, however, differential experience with pictorial representation, for example, may significantly influence the way the tasks are perceived, the specifics are selected, and the way such aspects are restructured. Some of you, indeed, may remember the non-language item in the Army Non-Language Test. The task was to cross out the picture that did not belong with the other four. Chinese inductees, invariably, viciously and erroneously crossed out the illustration of a rising sun, because of its symbolism for them.

Rulon's new semantic test should prove ultimately to be a fruitful lead. In essence, it sets the task of "learning" to associate a geometric symbol for a concept generalized from a number of drawings of worldly events. The process involves the acquisition of a symbolic glossary which is tested by requiring the subject to show his mastery of the glossary not only as individual signs but also in combined semantic and syntatic organization. Involved in the task of learning the glossary and in demonstrating mastery over it, is the additional one for the subject to infer what he is to do. Basically, the kind of learning is somewhat like associating a Chinese ideograph with a concept generalized from several pictures. In contrast with the more extensive spoken or visual vocabulary, the Rulon glossary approach involves very few signs, meanings, and syntactical patterns. Under such limitation, the process differs in complexity from the more usual tests of verbal intelligence. Rulon, indeed, finds that the correlation between Stanford-Binet mental ages and the score on the Semantic Test of Intelligence is very low for a constrained sample of feebleminded children. One reason, but not the only one, may be that the processes tapped by the Semantic Test are quite different from those appraised by the Stanford-Binet. The added evidence contrasting the relation of school achievement with the Semantic Test and with the Stanford-Binet supports the belief that the two tests are not measures of the same functions.

Test-makers apparently have tried to eliminate bias from the appraisal of intelligence by covering-up group differences, by eliminating tasks that make for group differences, or by substituting different processes in evaluating groups. Do such procedures really remove the bias from the measurement of intelligence? My answer is No. They do reduce, of a certainty, the amounts and kinds of information about test performance of separable groups. Scientifically, however, ignorance of difference is a costly way to produce unbiased tests of intelligence.

The objective psychologist cannot fail to see the *reductio ad absurdum* of making unbiased tests of intelligence. For instance, following the implications of Eells' procedure and findings, a test involving manipulation of numbers, geometric designs and stylized drawings will probably favor men and boys. Will it then be necessary to select from such items the few on which women and girls will be equal to men? And if this be accomplished, should only those items on which endomorphs make performances equivalent to ectomorphs be retained?

There can be little doubt that among some kinds of groups differences do exist. As a matter of fact, the wide range of general and specific tests of intelligence has made it possible to establish much of the available knowledge of differential psychology. Not only has the awareness of such differences led to the emergence of a more adequate understanding of the relative advantages and limitations of intelligence tests but it also has increased our appreciation of the significance of difference in the understanding of children as individuals, and in groups. In a democracy, such as ours, respect for difference as difference is necessary. There is no virtue in developing instruments so blunted that they decrease the amount of information. Perhaps the best method for reducing bias in tests of intelligence is to use them with the full knowledge that endowment interacting with opportunity produces a wide range of differences. Appraisal of the variation of different kinds of intellectual functioning requires many kinds of tests so that the differences can be utilized for the benefit of the individual and for the good of society. Intellectual functioning certainly does involve the ability to learn to adjust to the environment or to adapt the environment to individual needs and capacities by the process of solving problems either directly or incidentally. Such a concept recognizes a variety of different aptitudes for success with different kinds of problems. The full appreciation of the variety of aptitudes and the development of adequate methods for appraising them, should in the long run, ultimately lead to the production of enough information to eliminate bias.

As the psychologist develops tests to measure mastery of different contents and processes, he will obtain the evidence about the inequalities of opportunity for maximum development. With such information, the psychologist, in cooperation with educators and others interested in social amelioration, will try to make those social inventions which will allow all in our democracy to have an equal opportunity for maximum development of their potentialities. The full utilization of such social inventions and social engineering will not eliminate the established fact that there will be differences among individuals and between groups. When differences are reduced by the advantages of opportunity, the credit will be to the tests that showed their existence. Difference as difference is not bias, but the information about it will lead to the gradual disappearance of some kinds of bias.

Techniques for the Development of Unbiased Tests

PHILLIP J. RULON

A SEMANTIC TEST OF INTELLIGENCE

NON-VERBAL TESTS of intelligence have never been satisfactory. They have not correlated well with verbal tests of intelligence, nor with success in intellectual or academic endeavors.

In the case of many non-verbal tests, the intellectual operation called for does not seem to be the same as that called for in academic or intellectual pursuits.

The distinction between the usual verbal test and the usual non-verbal test is not so clear when easy items are considered, as when more subtle or difficult items are examined. The strictly non-verbal tests of the past have by and large retreated from the function they were trying to get at whenever they were made difficult enough to be useful in selecting a few of the more able members of the population tested.

The problem undertaken by the present investigators was to develop a testing technique which would be free from the more or less glaring shortcomings of the usual *non-verbal* test, and at the same time be free from some of the commoner defects of *verbal* tests.

The following defects in the typical *non-verbal* test were regarded as worthy of avoidance:

1. In the administration of some non-verbal tests, verbal instructions are employed to tell the examinee what is required of him.
2. The examinee is presented with novel material which deprives him of the opportunity to exhibit any use he may have made of opportunities to make ordinary observations of the surroundings in which he has lived.
3. A time limit is sometimes imposed which renders the non-verbal test a speed test rather than a power test.
4. Some non-verbal tests require the manipulation of concrete objects, such as blocks, marbles, or other simple familiar things. It is hard to contrive items making use of these materials such

that the items are difficult in the sense ordinarily understood by *intellectual* difficulty.

5. Some non-verbal tests require the reading of symbols (such as Arabic digits) which may be non-language strictly speaking, but which are nevertheless associated with the use of language in our culture.
6. Some non-verbal tests require a verbal response from the examinee.
7. Some non-verbal tests, such as form comparison tests, put a premium upon visual perception almost to the extent of rewarding visual acuity: the difference which the subject is required to detect between two geometrical figures may be so minute as to present essentially a problem of visual acuity.

The deficiencies in certain *verbal* tests which were regarded as particularly to be avoided include the following:

1. Some verbal tests give an advantage to persons from certain cultural backgrounds, regardless of the language employed. That is, the content is more familiar to persons from one culture than to those from another.
2. Some verbal tests require an exhibition of previously acquired knowledge, rather than testing a skill necessary for accomplishing a new task.
3. Some verbal tests are essentially speed tests.
4. Some verbal tests allow a free response which causes scoring difficulties.
5. Some verbal tests put a premium upon the examinee's facility with his native language.

It was the purpose of our work to derive a non-verbal test technique which would be acceptable on general grounds, and be free from as many as possible of these undesirable characteristics.

The work was conducted in the Harvard Graduate School of Education under a contract between the President and Fellows of Harvard College and the United States Government, represented by the Personnel Research Section of the Personnel Research and Procedures Branch of the Personnel Bureau of the Adjutant General's Office, Department of the Army.

The manner in which we have attacked this problem can be seen best, I think, by your now opening your test booklet to the left-hand inside page. This page pretty closely parallels the first page of our 42-page test booklet as it is now arranged. In giving the test, we make motions indicating that the symbol at the top goes with the five pictures. Motions then indicate to the examinee that the first symbol in the exercises is identical to that in the definition above. This is done without saying anything. Searching motions among the five options in the first exercise terminate in locating the COW WALKING in the third option. Motions of comparison between this picture and the fourth picture above and also motions of comparison between the symbol at the left and the symbol above terminate in the examiner's drawing a circle around the third octagon in the first exercise. Why don't you now draw a circle around the third octagon in the first exercise.

Similar motions of comparison terminate in the examiner's circling the fourth option in the second exercise. Suppose you now circle the JUMPING COW at that place. The motions are again repeated, still without saying anything, and the first octagon is circled in the third exercise. I suggest that you circle that STANDING COW and then go on with the rest of the page as our examinees are encouraged to do.

On the adjacent right-hand page you will find a lay-out very much like page 13 of our 42-page test booklet. You will see that the symbols at the left alternate between COW and JUMPING. For the first exercise we make motions indicating the similarity between the symbol in the exercise and the right-hand symbol above, and then make searching motions among the five options which terminate in the second option. Motions of comparison between this picture and the WOMAN JUMPING in the glossary above terminate in the examiner's drawing a large circle around the second option in the first exercise. In the second exercise motions of comparison are made between the symbol and the COW symbol above, and searching motions among the options terminate in the fifth option. Motions of comparison between this option and the WALKING COW above terminate in the examiner's drawing a large circle around the last option in item 2. Similarly in the next item the fourth option is circled by the examiner, after which the examinee is encouraged to circle the appropriate options on the rest of the page. Suppose all of you go ahead and do that at this time.

So far we have been engaged in a relatively simple intellectual operation which may be identified as a digit-symbol substitution exercise, except you may have noticed that in the second exercise on this page the WALKING COW was a mirror image of the one in the definition. Furthermore, in the third exercise, the JUMPING

CAT was not the same JUMPING CAT as in the glossary at the top. In the next exercise—that is the fourth one—you circled a jumping animal which was not shown at all in the glossary for JUMPING. In marking that exercise you must have abstracted the concept of JUMPING from the actions shown in the glossary at the top. You couldn't have marked that answer by a simple digit-symbol substitution.

On the back page of your booklet I have shown you what happens on page 24 of our 42-page booklet. In dealing with the first item, the examiner must here make two sets of motions of comparison. By such motions he shows that the first symbol agrees with the COW symbol above, and the second symbol agrees with the JUMPING symbol above. The searching motions among the options terminate with option 4. Then motions of comparison are used between this option and the JUMPING COW at the top left, and other motions of comparison between this picture and the JUMPING COW at the upper right. These motions terminate in the examiner's circling the JUMPING COW in the fourth option. In the next exercise similar motions of comparison terminate in the examiner's circling the first option. If you will circle this option, and in the next item circle the second option after comparing the symbols, I will then turn you loose on your own.

I have a few more remarks to make after all of you complete the exercises on this page.

As you may well suppose, the next step is to introduce three-symbol sentences, such as MAN BEATS HORSE or HORSE DRAGS BOY or BOY BEATS MAN. The highest level to which we are now going is to the four-symbol sentence, such as WOMAN KICKS DOG LYING DOWN or MAN BEATS WOMAN RUNNING, and the like.

I am sure you must have got the idea by this time why we called it the Semantic Test of Intelligence. What we have done is to imitate in a non-verbal test the semantic relationships presented in the typical low-level verbal intelligence test; that is, to require the subject to associate an arbitrary symbol with a worldly referant, to indicate his mastery of this association, and then to combine these symbols into groups in which the relationships between the symbols in each group are semantic or syntactical relationships.

In order to avoid putting a premium upon urban culture or amount of schooling, it was decided to use as worldly referants only the actors, verbs, and objects familiar in all western cultures, even the most primitive. These were felt to be sex differentiation, the young of the species, and domesticated animals, as far as the nominatives were concerned, and simple objects like bowls, stools, trees, etc.,—in addition to men, women, children, and common animals—for the objects of transitives. For intransitive verbs the most universal actions were used: standing, walking, running, jumping, sitting, and the like. For transitive verbs again the most primitive operations upon objectives were employed: pushing, dragging, lifting, beating, chasing, leading, etc.

The test is non-verbal to the extent of being administered without any word in any language being spoken by anyone.

The appearance of validity of the material is not merely superficial, since the operations required of the examinee are the simpler linguistic or semantic operations, not just operations thought up for the purpose of constructing a test. These operations are undoubtedly related to the operations of reading in any language.

It has been found possible to construct a test of substantial difficulty which does not seem to offer any reward for visual acuity or pure visual perception.

Also it seems possible now to produce such a test using such materials as not to give any advantage whatever to the Northern child over the Southern, the white over the colored, or the time-server in school over the bright youngster with less schooling.

S T I

Semantic Test of Intelligence

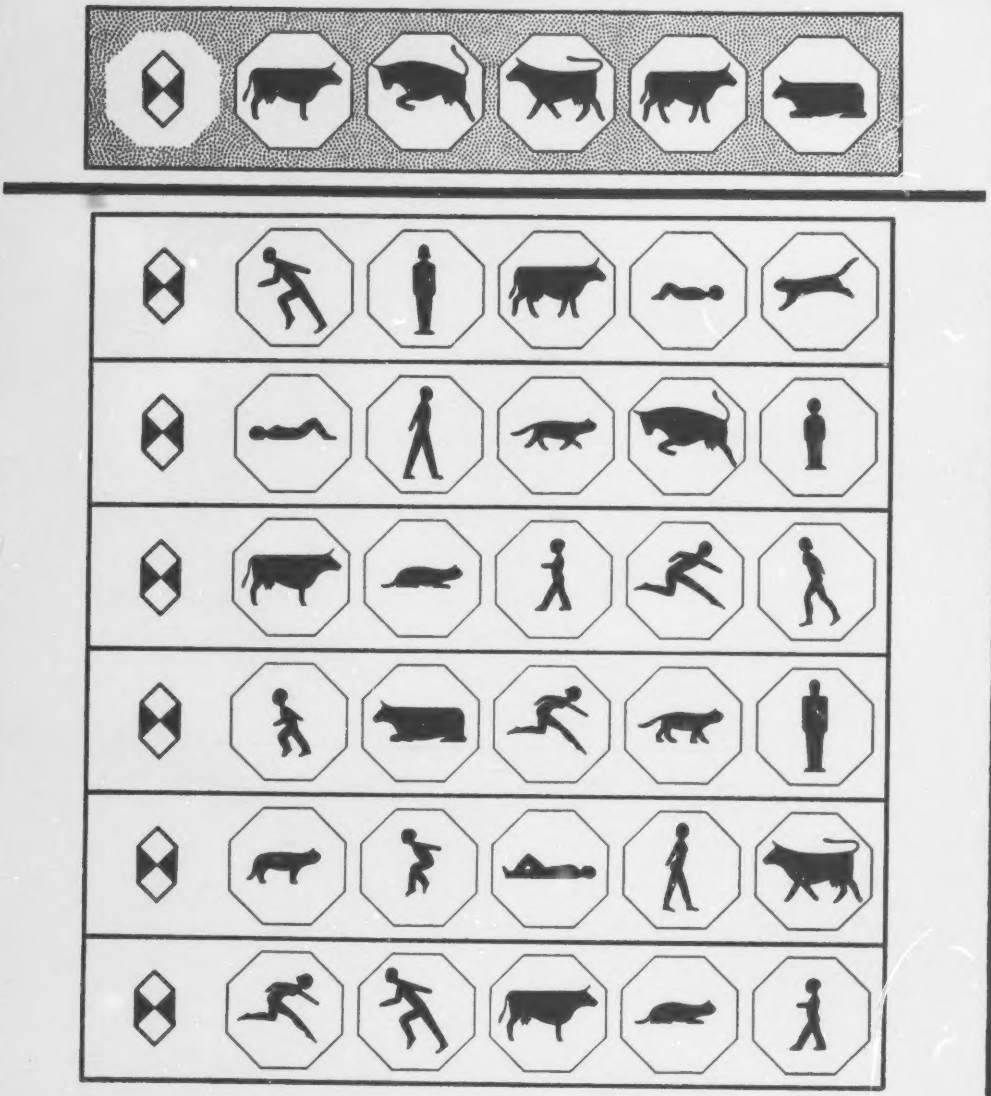
by

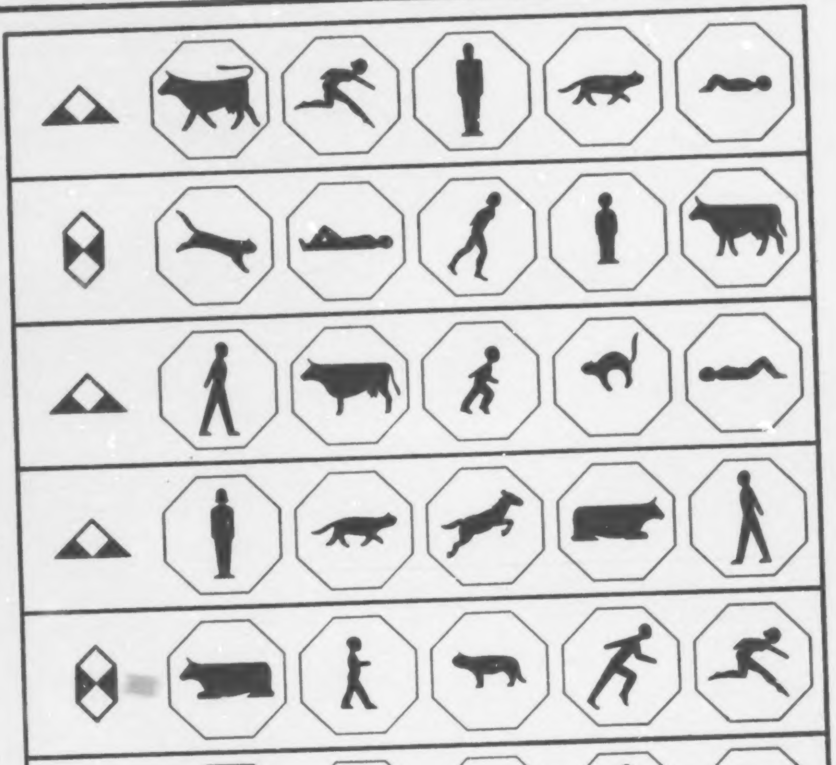
Phillip J. Rulon
Harvard Graduate School of Education

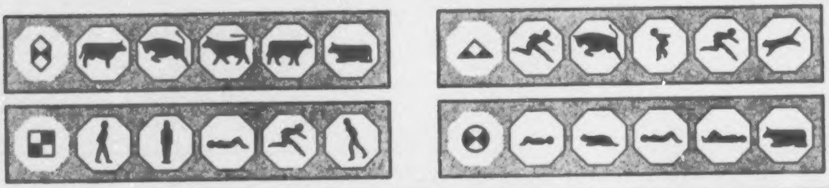
Form 00

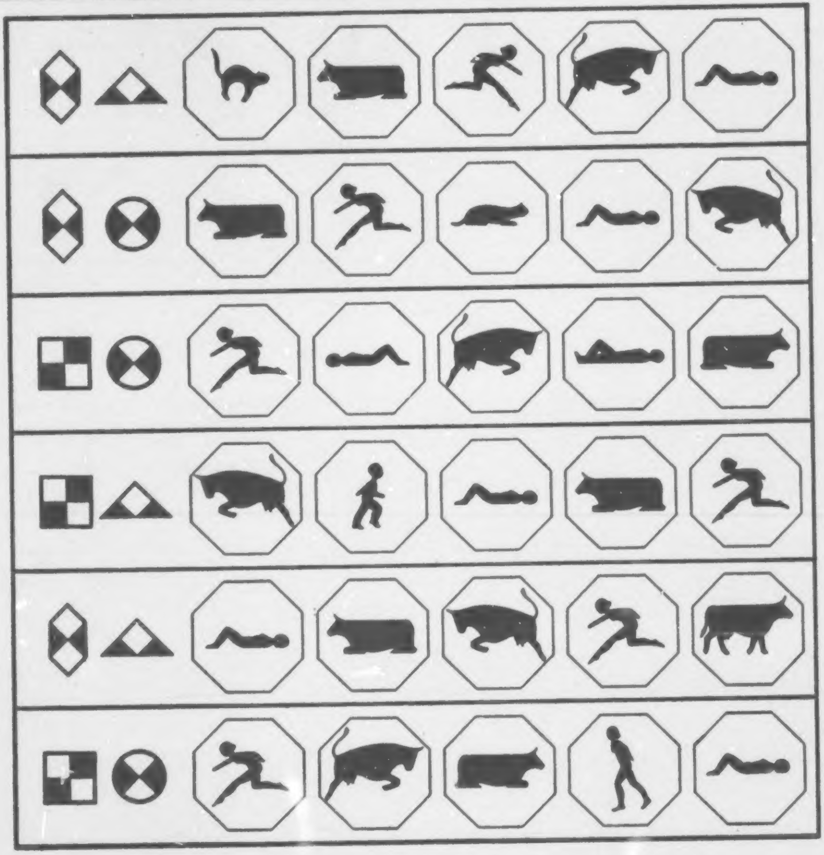
Special Edition for ETS Invitational Conference on Testing Problems
Saturday 1 November 1952

Copyright 1952
President and Fellows of Harvard College
Design Patent Applied For









Techniques for the Development of Unbiased Tests

ERNEST A. HAGGARD

This afternoon I will group my comments around four points:

1. What I believe to be the theoretical scientific conceptual model, or theoretical foundation, on which the test construction business has developed and thrived in America;
2. Some of the blind alleys and pitfalls we have been led into by the use of an inadequate and outmoded conceptual model;
3. Some specific research findings which bear on the topic of bias in intelligence tests; and
4. Some considerations which must be taken into account if we are to make any real progress toward a meaningful solution of the problem of developing "unbiased tests."

PART I

About three decades ago, the testing movement sank its tap root into the fields of education and psychology. This proved to be fertile soil since these groups, at the time, were first and foremost anxious to be scientific, objective, and quantitative. The influence of men like Thorndike and Watson was at high noon. Most of Thorndike's genius was devoted to activities which involved pioneering in areas where experimentation and quantification could be applied. Watson's recently formulated Behaviorism attempted to flush out of American psychology all of the subjective, conscious cognitive processes, and set up the dictum that only behaviors which could be observed objectively were worthy to be legitimate data for the science of psychology.

In these schemes, the subject was considered the equivalent of an independent, isolated, physicalistic machine; the experimenter imposed stimuli which in turn elicited responses. This machine could thus be manipulated and its characteristics studied by the psychologist, much as the physicist studied physical phenomena in his laboratory. The task of the scientist was to control (and measure) stimuli and other environmental conditions, to observe (and measure) responses, and to determine the relation between the two. The role of the scientist was that of a *deus ex machina*, equipped with an inelastic "foot-ruler."

Complex, adaptive behavior was seen, in this setting, as conglomerates of simple elements, whether they were called "neural connections" or "conditioned reflexes." Such were the psychologist's atoms. Given enough of them, and in the right proportions, you had what was called, for example, "intelligence" (cf. 37).

Under the influence of this philosophy of scientific method and procedure, the task of the intelligence tester was relatively easy—all he had to do was to present stimuli (items) to the subject, and determine the adequacy (some measure) of the subject's responses. But underneath all this was a more fundamental set of ideas, namely those which characterized the early nineteenth century physical sciences. These ideas were reflected directly or indirectly in the practices of the early, and to a large extent the present, testing movement.

Loosely stated, some of the underlying assumptions of this conceptual scheme which are relevant for our consideration today are: that the phenomena to be studied were stable; that they were a "closed system;" that the "closed system" functioned in a manner analogous to Newtonian formulations of thermodynamic laws; that the variables used in describing the phenomena were independent, and quantitative in linear, unidimensional terms; and that these variables or dimensions of behavior could be measured by the application of some form of external "foot-ruler," which could be applied by any trained impartial observer who was removed from, and was independent of, the phenomena being studied.

I will return to comment on some of these assumptions from time to time. But first, I want to say that it is too bad, and a little ironic, that the educators and psychologists who saw their path to scientific respectability in imitating the physical sciences, imitated a conceptual framework which was already discarded as inadequate by the very discipline which developed it. Clerk Maxwell had published in 1877 his *Matter and motion*, the work which opened a new era in physical science theory. But those educators and psychologists whose thinking and research were patterned after the classical model with which they were familiar, were intent on being "scientific" at any cost, and the cost has proved to be high.

PART II

What have been some of the effects on the field of intelligence testing that resulted, directly or indirectly, from following the theoretical model of classical physics? There have been several, but I will limit myself today to three major groupings; the confusion of problems with techniques; the confusion of facts with artifacts and the generation of pseudo issues; and the reluctance to consider approaches which deviate from orthodox theories and techniques.

1. *The confusion of problems with techniques.*

In a lucid discussion of this topic, Maslow (23) has listed several consequences that result when the techniques or means of investigation of scientific problems are confused with the problems themselves. Some of the consequences he lists are: the tendency to lay stress on elegance, polish, and technique, to over-value quantification indiscriminately and as an end in itself, to fit problems to techniques rather than vice versa, to develop an orthodoxy by those who use the proper techniques, which in turn tends to block the development of new methods, to exclude many problems from the jurisdiction of science, and to make scientists want to be "safe," rather than daring and creative.

I think that examples of all of these trends or tendencies can be found in the history of the testing movement. Perhaps it was because educators and psychologists felt so strongly the need to be "scientific," and at the time had precious little else that held out such promise, that they assumed that this would be a convenient escalator to scientific status. In any case, they seem to have devoted their energies to the development of the means or techniques, and have forgotten somewhat the basic problems they set out to solve. Indeed, in some cases, they seem to have substituted the means or techniques for the problems themselves. In making this switch, they were sometimes criticized that their tests did not measure intellectual potential after all; the reply has been that their tests did predict school achievement pretty well. They seem not to have questioned their purposes, but rather to have justified their techniques. But let us be more specific; let us consider the kinds of concerns that have preoccupied test constructors, with occasional illustrative references to our most esteemed test of intelligence, the 1937 Revision of the Stanford-Binet. I am selecting it, certainly not because it is more vulnerable to criticism than others, but because we probably know more about its standardization than we do about any other test, and because various other tests have used it as their criterion of "validity."

The primary concerns of most intelligence test constructors can most likely be summed up in three terms: item-difficulty, reliability, and validity—and probably in that order of importance. I have picked these three terms because they have formed the essential justification for many tests of intelligence. Personally, however, I don't think they are independent concepts at all, but I will try to speak of them separately, since they are supposed to be kept separate.

Let us take *item-difficulty* first. I believe that

> "in terms of the usual test-construction procedures, there is no absolute way of establishing the true item-difficulty, since the 'difficulty' of a given item, or set of items, is in practice defined 'operationally' by the proportion of children of a given chronological age who pass it.
>
> "This point may be clarified by tracing briefly some of the procedures used in test construction and standardization as follows: Let us assume that a test constructor finds it necessary to establish, first, that a given proportion (e.g., 50 per cent) of the persons with a given CA pass an item with a given sigma; and, second, that an item on an initial testing is found to be 'too easy.' The test constructor in this case usually makes the item 'harder' by either rescoring or rewriting it.*
>
> "In rewriting an item, two procedures are generally used to accomplish this desired purpose: either (a) to make the mental problem more difficult, so that more mental ability is required to solve it, or (b) to retain the same mental problem, but change the form of the item so that fewer children at a given age level pass it. In the latter case, this is done most easily—and most often—by manipulating verbal, etc., factors in the item, usually by using more 'difficult' (i.e., esoteric, unusual, or academic) vocabulary in presenting the mental problem-to-be-solved. And, because of the statistical definition of item-difficulty, it is not possible to tell whether the mental problem really was more difficult, or just accessible to fewer of the children in the standardization group because of the unfamiliarity of the vocabulary or other language forms" (17).

It is true that by this procedure the mean mental age is raised for the item, and that the sigma of the distribution may remain the same. But this type of standardization procedure leaves several rather basic questions unanswered. Let us consider two of them. First, we do not know what happened to the relative position of the individuals between the first and second distributions. Please bear with me while I make the following assumption: Just suppose that on the initial testing for our hypothetical item there was no difference between the performances of low-status and high-status children, and suppose also that there was a significant difference on the second testing. The shuffling of the relative position of individuals in the second case would not be apparent at all from the mean and sigma of the second distribution, arrived at by the procedure I have described, which is the usual one. Furthermore, I strongly suspect that in such cases, items are more often made "harder" than they are made "easier," probably because the floor of item-difficulty is set by the actual difficulty of the mental problem-to-be-solved, whereas the apparent ceiling of item-difficulty can be raised easily by the introduction of such artifacts as I have suggested.

---

* In group tests, which use a simple scoring method, the item is generally rewritten and again tested, or it is discarded. In individual tests where more complex scoring procedures are possible, rescoring as well as rewriting is used in the standardization procedure. In the 1937 Revision of the Stanford-Binet, for example, the tests and scores were revised six times in order to obtain "proper" distributions for items on Form L of this test (36, p. 23).

A second unanswered question shows how difficult it is for me to keep "item-difficulty" and "validity" separate. It has to do with the question of

> "whether it is necessary, or even desirable, to confound problem-difficulty with vocabulary-difficulty in intelligence test items. One reason for believing that these two aspects of an item often are confounded is that for many intelligence tests the vocabulary score usually has the highest correlation with the total test battery. (Terman and Merrill [36, P. 302], for example, cite a set of such correlations for single age groups which range from .65 to .91, with an average r of .81 for their test.) If problem- and vocabulary-difficulty are confounded (whether intentionally or unintentionally), it is highly unfortunate in view of the known differences in the extent to which children from widely different social classes are exposed to the academic language permeating most current intelligence tests. It seems apparent that the removal of a vocabulary-bias which favors middle-class children does not lower item or test validity, but indeed increases the validity if one attempts to measure problem-solving ability rather than vocabulary" (17).

The question of whether "vocabulary-bias which favors middle-class children" is present in an item cannot be answered by the statistician; it can only be determined on the basis of socio-anthropological field research.

Next, *reliability*.* I have difficulty in understanding clearly what is meant by reliability—perhaps because there are so many types of it. But I do get the impression that it is almost a rule-of-thumb to say, "If your test-retest reliability leaves you with a large error showing, try the split-half method, and then increase the reliability a little more by using the Spearman-Brown formula."

---

* Besides being used as a measure of the stability of some phenomenon or behavioral characteristic over a period of time—which I have discussed here—the term "reliability" is often used in other, and quite different, senses. These include the homogeneity of the measures of some particular characteristic within a person or situation which are sampled at a given time; and as the index of the consistency (or "objectivity") of the scoring of a particular sample of behavior by two or more persons or methods of evaluation.

I am not entirely facetious in making this statement, because last year we obtained both split-half and test-retest reliabilities on a well-known intelligence test. The uncorrected split-half reliability was .97; the test-retest reliability, after about sixteen months, was .67. (The number of cases was 68.) I will not mention the particular test, because I suspect this sort of thing happens with more than just this one, and besides, the general problem, and not the specific case, is the important thing for us to consider. In this connection, I would like to mention Gulliksen's point in Ch. 17 of his book on the *Theory of Mental Tests* (16), namely that the use of split-half reliability on speed tests, where the unanswered items are counted as being incorrect, yields spuriously high reliability coefficients. The Stanford-Binet test is not open to this criticism, but several others are.

In the sense that reliability is commonly used, what does it really mean? It makes sense to me only in terms of the classical physical science model, one assumption of which was that the object or phenomenon under investigation remained stable. Thus, if you measured a lead brick with a given "foot-ruler" at one time, and if you went back on a later occasion and measured it again, and if your "foot-ruler" gave you the same reading the second time, you could say it was reliable; your measuring instrument did not change. You assumed all along that what you were measuring did not change. But the problems of measurement that we have to face are not only much more complex; they are in fact different. In a sense, for us, the experimenter, the test items, and the subject cannot be clearly separated. The person who administers and scores the test must be thought of as part of the "foot-ruler," not just the particular test alone.

The use of the classical model in attempting to measure intelligence, for example, does not fit the facts. If we try to use this model, we would have to say that many conditions (such as the subject's motivations, past experiences, attitude toward the examiner, the test situation, and a host of others) change so drastically "the size and shape of both the object of measurement and of the foot-ruler" that these latter terms cease to have meaning. Gulliksen referred to this problem when he said that "a significant contribution to item analysis theory would be the discovery of item parameters that remained relatively stable as the item analysis group changed; or the discovery of a law relating the changes in item parameters to changes in the group" (16, p. 392).

Rather than trying to make the old system of measurement work, it might be better if we would go back and examine our basic measurement assumptions, and modify them and our technical procedures to fit the requirements of the total measurement situation as we know it to be.

Finally, *Validity*. The concept of validity has sometimes been treated even more casually than reliability. In looking through *Measuring Intelligence* (36), I found a reference to the fact that for the 1937 Revision of the Stanford-Binet, items were selected "that experience had shown to yield high correlations with acceptable measures of intelligence" (p. 7). I was unable to find a clear statement of the "acceptable measures of intelligence." However, Terman and Merrill (36 p. 9) considered validity to be of primary importance in selecting test items. Validity, they said, was judged by two criteria:

> (1) "Increase in the percents passing from one age (or mental age) to the next, and (2) a weight based on the ratio of the difference to the standard error of the difference between the mean age (or mental age) of subjects passing the test and of subjects failing it. The use of such a weighting scheme was prompted by the obvious advantage of being able to utilize the data for all of the subjects who were tested with a given item" (p. 9).

These authors go on to point out that

> "Increase in percents passing at successive chronological ages is indirect but not conclusive evidence of validity. Height, for example, increases with age, but is known to be practically uncorrelated with brightness. Increase in percents passing by mental age is better, but exclusive reliance upon this technique predetermines that the scale based upon this criterion will measure approximately the same functions as that used in selecting the mental age groups" (p. 10).

This seems a rather scanty justification of validity, in view of the great expenditure of time and energy that went into the standardization of this test. McNemar later threw some light on the "validity" of the Stanford-Binet. He said that "the ultimate criterion of validity was correlation with mental age or its equivalent in point score on the composite of the two scales" (25, p. 4). In his *Psychological statistics*, however, he says that "the correlation between two determinations is

. . . termed the *reliability coefficient*" (26, p. 128). From such statements it is not clear to me whether a measure of validity, or a measure of reliability, was used in standardizing this test.

But let us return to the Stanford-Binet. Since the 1937 Revision was validated in terms of the 1916 Revision (25, p. 13), I read Terman's *The Measurement of Intelligence* (33). Nowhere did I find any clear statement of his validating procedure. Terman's assertions that "the validity of the I.Q. as an expression of a child's intelligence status . . . follows necessarily from the similar distributions at the various ages" (33, p. 68) did not entirely satisfy me, and his statement that "a test which makes a good showing on this criterion of agreement with the scale as a whole becomes immune to theoretical criticisms. Whatever it appears to be from mere inspection, it is a real measure of intelligence" (33, p. 77) left me with the feeling that Terman was really making a case for *"faith validity."* However, he did report the correlation between the I.Q. and teachers' estimates of the children's intelligence. It was .48, which "is both high enough and low enough to be significant. That it is moderately high in so far corroborates the tests. That it is not higher means that either the teachers or the tests have made a good many mistakes. When the data were searched for evidence on this point, it was found . . . that the fault was plainly on the part of the teachers" (33, p. 75). The correlation between I.Q. and school success was given in another source (34, pp. 104-6) as being .45. This correlation is not startlingly high, but in view of the great amount of rather prosaic rote learning and recitation required in our present public school curricula, I do not think it would be very flattering to a test which purports to measure an individual's complex, adaptive, higher mental abilities if it correlated too highly with school success.

There is another point that I would like to mention with regard to the standardization of the Stanford-Binet. We are told that in both the 1916 Revision (33, p. 52) and the 1937 Revision (36, p. 15) "schools of average social status were selected in each community." As we know, this means middle-class schools, which are usually attended by middle-class children. This is a rather serious sampling error, since research has shown that the concomitants of social class, the range of experiences, motivations, etc., in turn influence performance on our present intelligence tests (e.g., 4, 8, 9, 13, 17, 19, 22, 32, 40). It is analogous to the error that we would make if we wanted to determine the social and emotional behavior of individuals of all ages up to maturity, and took an "average" age sample, namely adolescents. It does not follow that they would be truly representative of either young children or adults. The use of "average schools" would have been a good sampling short-cut *if* the factors which influence intelligence test performance were randomly distributed along a linear continuum of social status. Again, whether this is a correct assumption cannot be answered by the statistician; such decisions must be based on socio-anthropological field work among various social-status groups in our society.

A further sampling problem in connection with the 1937 Revision is seen in the fact that Terman and Merrill selected their standardization groups from white, native-born children, and on the basis of census norms for *employed males in 1930* (36, p. 14). Even for this biased criterion, they selected too many children from high-status and too few from low-status families. The extent of the discrepancy for the seven classifications used is indicated by a chi-square greater than 500. Because of such sources of bias, W. L. Warner (38), on the basis of his research on the differences in cultural behavior and experience patterns of various social class groups, estimates that the Stanford-Binet should, on these grounds, be appropriate for testing fifty, or perhaps even sixty-five, per cent of the children in our population.

2. *The confusion of facts with artifacts and the generation of pseudo issues.*

The history of intelligence testing in America has been fraught with a series of violently contested "issues." I would like to make some comments about the so-called "constancy of the I.Q."—although I might also have chosen to speak of the "nature-nurture controversy." When some people speak of I.Q. constancy, they assume that the person, because of his genetic inheritance, is born with a given level of intellectual potential, and for better or worse, it is his for life. It is even more constant than, say, one's hair because one can dye, curl, or even lose his hair—but the I.Q. remains faithful to the bitter end. Terman expressed this position in speaking of gifted children thus: "Their high I.Q. is only an index of their extraordinary cerebral endowment. This endowment is for life. There is not the remotest probability that any of these children will deteriorate to the average level of intelligence with the onset of maturity" (33, pp. 102-3). It is this intellectual "something" that is said to be measured by intelligence tests.

What is the evidence for this belief, even though many investigators have found, or some investigators have found many times, that children often do receive the same I.Q. within rather narrow limits, when tested from time to time. That is to say, when groups are retested after from two to five years, one can be reasonably sure that their I.Q.'s will vary not more than about five points in either direction from their previous score. Buttressed with such findings, it has apparently been assumed that "the constancy of the I.Q. is the main argument for its being determined by heredity" (19, p. 302).

Now, let us look behind these "findings" for a moment, and how they were obtained. If we test and retest average middle-class children from classrooms, we can be almost certain that they come from families that are stable in the neighborhood, maintain a certain style of life, and that the over-all parental values, their systems of rewards and punishments, their expectancies for their children, the type and range of experiences of the child, the strength of his desire to do well in school, and so on—that these conditions will remain about the same from test to retest. Thus, even if environmental factors do influence intelligence test performance, we would not be able to observe their effects under such conditions, since such effects would be held relatively constant. Some theorists remind me, by their logic, of the man who thought his thermometer was stuck because it always gave the same reading, even though he kept his house at the same temperature. However, in his discussion of occupational differences in obtained I.Q.'s on the 1937 Revision, McNemar's point is well taken that "a quarter of a century ago such an accumulation of data as we can here present would have been hailed as definite proof that intellectual differences have an hereditary basis, but at the present time these data will not be regarded as of crucial significance in a field of controversy" (25, p. 35).

In short, I am suggesting that the "constancy" may have been an artifact of how we obtained our data, and that unwarranted generalizations from such data generated an "issue" over which we spent too much adrenalin, time, and energy. There are other "issues" that would probably fall into the same class, and social status is one of them. I think that such controversies would cease to be of central importance if we knew more about the nature of the phenomena we are attempting to measure, if we were better able to formulate and appreciate the relevant parameters that concern what we mean by "problem-solving ability," and if we could learn ways to cut through various aspects of the testing situation which are actually irrelevant to our basic measurement purposes, but which can contaminate our test items and standardization procedures, and sometimes have.

Earlier in this paper I questioned the long-range value of trying to be too "scientific" too soon, of setting out to develop tests that are justified primarily in terms of their statistical characteristics, and that may correlate with something. I have also stated my belief that the test *per se* is only a part of the measurement situation with which we must concern ourselves. These points will be touched upon later in this paper.

3. *The reluctance to consider approaches which deviate from orthodox theories and techniques.*

This tendency is of importance only in so far as it serves to impede scientific progress in the development of new ideas and knowledges. From time to time, new approaches to basic problems, or a questioning of the established "facts" in a field, meet with rebuke or censorship. Sometimes they are justified and sometimes not, but this tendency occurs in every field (cf., 39, ch. 4).

In this connection, I decided last week to check the reviews of the book by Eells and others, *Intelligence and cultural differences* (13). I was able to locate ten reviews in general scientific, psychological, and sociological journals. It was apparent at once that all of the reviews except those in the field of psychology were either mere factual reporting of the research and ideas, or very favorable, with such lauditory statements as "the book might well serve as a model for social science research" (24, p. 45), or "this very important study will be of great interest to psychologists as well as to social scientists, particularly to those concerned with constructing and giving tests" (30, p. 209). The reviews in psychological journals were somewhat less enthusiastic.

Could it be that our colleagues in other disciplines, such as sociology, are not sufficiently familiar with the problems of this field, or sufficiently knowledgeable to judge adequately the value of such work? Perhaps. Of the reviews of this book in psychological journals that I have seen, McNemar's criticisms were the most just, although his praise for its value was barely audible. McNemar closed his review (27) with the statement, "Eells, perhaps in tune with his mentors, concludes that 'variations in opportunity for familiarity with specific cultural words, objects, or processes, required for answering the test items seem . . . to be the most adequate general explanation for most of the findings'" (p. 371). I wish McNemar had seen fit to cite the following paragraph, also on p. 68 of this book (13), which reads,

> "It seems likely that status differences in response to intelligence-test items are not due solely to any single simple cause but are the result of various types of factors, quite possibly including both genetic or developmental differences in real ability, on the one hand, and motivational and culture-bias factors in the tests, on the other. Interpretation of I.Q. differences between pupils of differing cultural backgrounds should, therefore, be made with extreme caution."

or the statement on p. 357, which reads,

> "Another important finding of the analysis reported in this chapter is the rather substantial number of items showing large status differences for which no reasonable explanation can be seen. . . . The presence of such a large proportion of unexplained differences should, however, lead to caution in accepting the idea that all status differences on test items can be readily accounted for in terms of the cultural bias of their content."

I point this out because many persons have mistakenly assumed that the people at the University of Chicago believe that the influence of heredity is not reflected in intelligence test scores. No one at Chicago ever said that, but rather that our present tests "measure" a very great deal besides hereditary potential.

Another instance of reluctance to accept the "Chicago Studies" came to my attention recently when I learned that a research report from there was not accepted by the editor of a well-known psychological publication. I was a little taken aback to read, among other things, the editor's comment as follows: "I guess that that's something that troubles me most: the fact that the implications of this study are so grossly different from what is generally believed." This comment, although perhaps more revealing, is hardly more encouraging to progress than the one of an editor who allegedly said, "Here is your paper, somebody wrote on it." However, such a finger-in-the-dyke approach is futile, especially when the main stream of scientific thought and methodology has long since gone in another direction.

PART III

It would be impossible to attempt here a survey of the mass of research findings which bear directly or indirectly on the topic of bias in our current intelligence tests. Such a survey would have to draw on materials from such fields, for example, as sociology, anthropology, psychology, psychoanalysis, and education. It would have to deal with the host of factors that touch on the broad problem of how people come to behave the way they do, and why. I will limit myself here to only a token sample of the kinds of findings in the psychological literature that are obviously relevant to our problem. These include the relation of the individual's early learnings to his later behavior, especially as they effect intelligence test performance (19, Chs. 6, 7, 11), the role of learning to learn in effective problem-solving behavior (18), and the role of emotional or personality disturbances, especially as they result in intellectual malfunctioning (40).

I would, however, like to discuss some research findings which bear more directly on the problem of bias in intelligence tests. This experiment is a part of a larger research program that has been going at the University of Chicago for seven years, under the leadership of Allison Davis. The experiment I will discuss was designed to investigate experimentally some of the many factors which are known to be culturally determined, and which influence the performance of children on our present intelligence tests. The factors which, it was felt, could be studied realistically and controlled experimentally are formulated in terms of the following experimental conditions: (a) social-status, (b) practice, (c) motivation, (d) the form of the test items, and (e) the manner of presentation of the test items (17).

You have already been given a brief description of the major variables and experimental conditions, the matching variables, the control variables, and how the data were analyzed (See Appendix). To save time, I will go directly to a summary of some of the major findings of this experiment. They are as follows:

1. "The condition of Practice facilitated the gain in performance of the high-status children who took the Standard form of the Retest, and the gain of the low-status children who took the Revised Retest.
2. "The condition of Motivated Practice interfered with the gain in performance of both groups of children who took the Standard Retest; this was especially true for the high-status children.
3. "The low-status children, when motivated, did significantly better on the Standard Retest than the low-status children not thus motivated.
4. "Children from both social-status groups made much greater gains on the Revised, as opposed to the Standard form of the Retest, with the low-status children showing the greater gain.
5. "Some item-types (e.g., analogies, opposites, classification) can be revised more easily than others (e.g., syllogisms) to reduce 'mid-

dle-class bias.'

6. "Children from both social-status groups performed better on a Revised Initial Test than on the Standard-type Initial Test.
7. "High-status children showed a slightly greater gain when they took the Revised Retest in the traditional (Silent) manner, whereas the low-status children showed an additional gain in performance when the Revised Test was also read aloud to them.
8. "The Initial Test and Retest of 40 items were not given under strong pressure of time. Many more children from both social-status groups passed the test items than one would expect from the standardization norms for these items.
9. "Even though the various experimental treatments and conditions influenced the retest scores of children in the two social-status groups differentially, when the effects of all such treatments and conditions were thrown together, there was no significant difference between the two groups of children in their ability to learn to solve intelligence test problems.
10. "Children from both social-status groups showed greater gain in performance when tested on tasks and under conditions which were relatively more familiar to them.
11. "The mere revision of the test items was not in itself sufficient to reduce the difference in performance between the high-status and low-status children. The marked discrepancy between the two groups was only decreased when the conditions of Motivation and Practice were also present—that is to say, when there was also a decrease between the two social-status groups in the difference in their familiarity with, and motivation to do well on, the test items.
12. "All of the statistically significant differences attributable to the conditions of Practice, Motivated Practice, and Motivated Retest occurred in connection with the Standard Retest. The Revised Retest was not so influenced by such conditions, which are essentially irrelevant to the measurement of mental ability, but which are determined in large part by the concomitants of social-status" (17).

PART IV

The mere existence of this panel on unbiased tests implies some interest in the measurement of potential (e.g., for abstract reasoning or problem-solving ability), and some dissatisfaction with tests whose essential justification is in terms of some criterion of expediency.

Furthermore, by potential, I assume we mean first, an individual's potential at the time of testing, in terms of his theoretically maximal ability to perform, rather than some hypothetical innate, genetic potential, and second, that the role of social-status is one of the factors that interfere with our precise evaluation of present potential.

There are two broad levels on which we may approach the solution of the problem of possible bias in tests which attempt to measure, for example, intelligence. One is the theoretical level, or how we conceptualize and formulate our research problems. The other is the technical level, or what specific knowledges we should acquire, and what steps we can and should take in attacking our problems. I will discuss each of these levels briefly.

The Level of Theory. From time to time this afternoon I have made the classical physical science conceptual model out to be a scapegoat, the source of all our ills. Some of you may say I have carried this point too far—and I would agree with you. A more accurate statement would be that if we had been more articulate about some of the assumptions we have unwittingly made, we would have ceased to make them a long time ago. I really think our chief weakness has been that we assumed we were being "scientists" because we performed some of the scientific rituals, and we assumed our "facts" were valid because they were stated in quantitative terms. Others of you may say that I have been unjustified in spending so much time talking about vague "theoretical conceptual schemes," that we have work to do, so let's get to it. The best reply I know to this position was made by Einstein and Infeld, who said that "the formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from new angles, requires creative imagination and makes a real advance in science" (14, p. 95).

Frankly, I do not know of any conceptual model which is fully articulated and appropriate for the field of testing. But I think it is clear that the system we have been using is quite inadequate and inappropriate, as I have tried to point out from time to time this afternoon. Perhaps I can be more explicit by the use of an example which is admittedly exaggerated. It is a case cited by Boas (5) of "a certain psychologist who asked a native for the name of his mother. On receiving the answer 'Whom do you mean?' he marked intelligence as zero, because the man did not know his own mother. The psychologist did not know that the mother, her sisters, and all her parallel cousins are designated in the native language by a single term, and the situation did not make it clear that the own mother was meant" (p. 14). This example differs only in degree, I believe, from many practices that have existed in the administration and interpretation of intelligence tests.

In concluding this section I would like to say that since the time of Lobachewsky in geometry, Boole in algebra, Maxwell in physics, and Spemann and Weiss in biology, more flexible, general, and useful theoretical systems have been developed. In looking through recent issues of the A. A. A. S. publication *Science,* I found a number of papers which may give some helpful direction to our thinking in this field. Some have to do with applications to such fields as physics (6, 28), and genetics (12). But papers which I believe have rather direct relevance to some of the problems that we are confronted with include von Bertalanffy's discussion of the concept of open systems in biology (3), Bentley's use of the transactional approach in the general theory of inquiry (2), and a series of three papers by Cantril, Ames, Hastorf, and Ittelson on psychology and scientific research (7). Such approaches seem to me worthy of our careful consideration, in the hope that they may help us better to formulate our problems, and look for (and perhaps find) more fundamental solutions to them.

*The Level of Practice.** What do we mean by bias? In a general sense, whenever we speak of bias we refer to the influence on our test scores of factors irrelevant to the purpose of our measurement, and which can change any of the moments of our score distribution. The degree of bias and purity of measure, or validity, are inversely related.

Various suggestions have been made for developing unbiased tests. Binet (4), the pioneer in this field, was the first to point out that tests of intelligence should be free from the influences of various knowledges, skills, language usages, and other aptitudes which result from specific training in the home or school. In attempting to develop a test which was not contaminated by such experiential influences, Binet, early in his career, attempted to eliminate such culturally biased tasks from his test battery.

---

* I am aware that it is possible to make a case for tests (a) which are justified in terms of some criterion of expendiency such as prediction of school success, or (b) which are limited in their applicability to only a segment of the population, such as urban or middle-class groups. One certainly can argue for the existence of such tests, but in doing so I believe that the test constructor and the test publisher are obligated to make explicit and public answers to such questions as the following: Is this sufficient justification for the existence of the "intelligence test," since previous grade average would probably predict future school success as well as, or perhaps better than, such a test? Are the consumers of such a test led to believe that it is a measure of intellectual potential or problem-solving ability, and so act upon this belief (cf. 35)? Will such a test be limited in its use to the groups for which it is appropriate, and if not, will the misapplication of such a test serve to perpetuate the current wastage of large reservoirs of intellectual potential in our society?

In this country, Thorndike and others (37) set forth a few principles to serve as guideposts. First, they suggested that "intellect is the ability to learn, and that our estimates of it are or should be estimates of ability to learn To be able to learn harder things or to be able to learn the same thing more quickly would then be the single basis of valuation" (p. 17). In terms of constructing intelligence tests, they suggested that "the wisest procedure at present is to equalize environmental forces by using a wide variety of data with which all individuals have had adequate experience" (p. 462). Other suggestions included the use of novel tasks, "so that at least no person will have been taught to do that particular task by environmental forces" (p. 437), and the use of "tasks that are so familiar that everybody has had somewhat nearly adequate environmental stimulation to master them" (p. 439).

Recently, Davis (e.g., 8, 9, 13), largely on the basis of extensive research in this field, has also dealt with this problem at some length, and has reaffirmed and extended the early position of Binet. Davis (9) states that:

> "The crucial problem raised by the attempt to compare scientifically the capacity of any two individuals to learn is that of finding situations with which the two individuals have had equal experience. To state this issue more exactly, two major systems of behavior are involved in problem-solving. They are (a) the individual's genetic equipment for problem-solving; and (b) the individual's particular cultural experience, training, and motivation, which have developed certain areas of his mental behavior and certain skills more than others. In a test of general hereditary capacity, the second factor must be equalized for all those tested" (p. 301).

It will be noted that Davis has repeatedly emphasized that the condition of "equality" must be expanded to include such considerations as the manner in which the test is presented to the child, his attitude toward the testing situation, and his motivation to do well on such tests—as well as equality of experience in relation to the form and content of the problems used in the test. In constructing their *Test of General Intelligence*, Davis and Eells used test problems which are "(1) taken from the major areas of children's experience and (2) which

106

are not likely to have been previously taught in home or school" (10).

How can we approximate the various types of "equality" necessary to establish a minimal degree of bias in our tests? To answer this question, it is important first to consider the extent to which we can identify a possible bias variable. Some variables, such as whether a person is male or female, are easy to identify; others, such as social class or experiential background, are not so easy to identify. In the former case, if we want a test which is not biased in favor of one sex, our task is easy. We can, first, retain only the items which do not show a sex difference statistically, or second, we can balance our items, so that each sex is favored equally, or third, we can measure mental processes which are independent of sex differences. Since the first and third possibilities differ most sharply from a methodological point of view, let us consider them. In the case of a variable such as sex, these possibilities ultimately achieve essentially the same result. The test, and all its items show no bias in favor of one sex over the other.

It is important to note, however, that this equivalence of final result does not necessarily hold when we are unable to identify the bias variable. Under such circumstances we cannot say with confidence that the first method will give us the same result as the third method. For example, if the test constructor is unaware of the existence of a variable such as social class, or has no way of measuring it, or ignores it, his statistical procedures will provide no safeguard against its entry as a source of bias in his test. (The same argument would apply to such possible bias variables as ethnic background or rural-urban differences.) As a matter of fact, as I suggested earlier, such bias has been known to turn up in the guise of "empirical findings."

I shall conclude my discussion with a consideration of two questions: First, what do we know about the presence or absence of bias (or "equivalence" as I have used the term) in our present intelligence tests; and second, what are some of the considerations that must be taken into account if we are to minimize bias (or maximize "equality") in future tests of mental ability?

1. *Sources of bias in present tests.* There has been remarkably little basic research in this area, in spite of a recognition of the importance of the problem.* On the basis of various studies (e.g., 13, 17), however,

---

* Terman, for example, observed that "age and environment may effect almost every test to a greater or lesser degree. To determine the exact extent to which this may be true for even a single test, would require an extensive investigation" (34, p. 135).

we are able to say with some confidence that, by and large,

a. Our present standard tests of intelligence do not really meet any of the above criteria of "equality," but rather they are highly subject to such culturally determined factors as degree and type of previous experience with the content, language usages, etc., of our present tests, as well as the child's motivation to do well on them.

b. A very large proportion of the item types characteristically found in our present intelligence tests cannot be made to demonstrate "equality" by a verbal face-lifting. The academic nature of problems and their content, as well as the manner in which the problems are presented, are sufficiently artificial to preclude their use in tests which are unbiased for large subgroups in our society.

2. *Some considerations for the development of unbiased tests.* In setting out to develop unbiased tests, we can be certain that there is no simple set of "techniques" or any rule-of-thumb approach to the problem. Actually, our task is complicated by the fact that there are certain aspects of our problem about which we can do nothing, but which are potential sources of bias in our tests. These include the totality of experiential and cultural heritage which the child brings to the testing situation, and which may range from possible functional deficiencies resulting from early nutritional deprivations to specific training on various types of tasks found in our tests. In any case, it is safe to assume that "differences of early experience can produce differences in adult problem-solving that further experience does not erase" (19, p. 299).

But our task is by no means hopeless, because there are a great many aspects of our problem that we can do something about. For all practical purposes, I believe that a good point of departure is to reevaluate the following aspects of the testing situation for their possible contribution to bias in tests of mental ability: (a) the construction and standardization of the test; (b) the nature of the mental processes measured, and their relation to effective behavior; (c) the attitudes, value systems, motivations, etc. of the persons taking the test; and (d) the manner in which the test is presented. These aspects of the testing situation cannot, of course, be clearly separated, but for the sake of this discussion I shall attempt such an artificial division.

a. The construction and standardization of the test. By and large, we can be certain that the test constructor is a middle-class individual, being a professional person with college training. His habits of thought, language usages, etc. reflect his middle-class culture, and when he writes test items, they too are likely to reflect this background. The way for him to avoid ivory-tower item writing is to learn, on the basis of research, as much as he can about how other sub-cultural groups in our society live, the words they use, their meanings, etc., and then to write items which do not favor one sub-group more than another.

Earlier in this paper, in discussing such topics as validity and item difficulty, I already considered possible sources of bias that might arise in the standardization of intelligence tests.

b. The nature of the mental processes measured, and their relation to effective behavior. It is not always clear just what mental processes are measured by "intelligence tests," or whether the same processes are measured for various age levels (cf. 20).

Some research findings bear on this question. In one study (1), children were asked to give the reasons for their answers to intelligence test items. In the case of one analogy item, 35 of the 60 children tested marked the "correct" response, but not one of these children gave the "correct" reason for marking it. The reasons given were on the basis of rhyming, synonym, etc., but not on the basis of making the analogy—the process which the test constructor assumed was being measured.

In another study (11), the test constructor wrote out the mental processes he thought were being measured by the items in his published test. It was found that for some items over fifty per cent of the 152 nine- and ten-year old children gave logically defensible reasons for marking answers considered "incorrect" by the test constructor. Furthermore, whenever more than one logically defensible answer to an item was given, the middle-class children tended to give the "correct" answer (in the opinion of the test constructor), whereas lower-class children tended to give the "incorrect" answer (in the opinion of the test constructor).

Perhaps an even more fundamental question has to do with whether the mental processes purportedly measured in intelligence tests bear a close relation to intelligent, effective behavior in life situations. Boas (5) defined the intelligence of a people in terms of "their ability to adapt themselves adequately to the problems of their life" (p. 11). In a general sense this appears to be a defensible position. But if test items are selected primarily in terms of certain statistical criteria,* it is not certain that such tests will predict intelligent behavior in this more general sense.

With regard to this problem, Davis and Eells (10) take the position that

> "In real life, the types of mental problems which the individual actually meets can seldom be solved by reference to specific instructions or memorized formulas . . . the individual has to learn how to organize his own data, to learn how to define the problem, and to learn how to develop a method for solving the problems as defined. He is 'on his own'; he has to find a way to solve the problems."

Consequently,

> "An intelligence test should approximate these conditions as nearly as possible. The test should be designed to measure what an individual can do in solving mental problems similar to those which arise in his general experience."

c. The attitudes, value systems, motivations, etc. of the persons taking the test. It is clear that many groups in our society appraise the testing situation differently from middle-class children, and may feel intimidated by, or not motivated to do well on, our present intelligence tests. It is also clear that such "non-intellective" factors as rapport, attitude, and motivation substantially influence performance on intelligence tests (e.g., 4, 8, 15, 17, 21, 29, 40.) Since the child's prior attitudes and motivations cannot be changed at the time of testing, then perhaps the testing situation can be made minimally threatening and maximally motivating to all children. This would necessitate both the construction of tests which are, in themselves, maximally interesting and motivating to all children, and the creation of a favorable "atmosphere" in which to give the tests. To use problems which are meaningful to all the children being tested would also serve to stimulate them to use their problem-solving ability to a maximum in the testing situation.

---

* Terman (33) states that he eliminated certain tests from his battery "which have been considered excellent," because they "proved to be so little correlated with intelligence that they had to be discarded" (p. 56). He defined intelligence here in terms of the score achieved on the total scale.

d. The manner in which the test is presented. If we are to develop unbiased tests of intelligence or problem-solving ability, it seems clear that the tests should be so presented that all the children tested have an opportunity to understand equally the problems-to-be-solved, so that they can utilize their problem-solving ability when taking the test.

One aspect of this question has to do with the academic vocabulary often used in presenting the test items. In a study based on 511 cases (32), it was found that when the words used in standard intelligence tests were made into vocabulary tests, from two-thirds to three-fourths of the terms were better known ($P = .05$) by middle-class than by lower-class children. Such a source of bias could easily be removed by presenting the test problems in terms which are equal in familiarity and meaning to all children taking the test.

Another important aspect of test presentation has to do with the emphasis placed on speed in many of our intelligence tests. In this connection it has been pointed out that

> "Speed is influenced both by cultural attitudes concerning the importance or unimportance of speed, and also by personality and motivational factors, such as competitiveness, conscientiousness, compulsiveness, exhibitionism, and anxiety" (10).

It was found in an experiment reported earlier (17) that many more children from both high- and low-status groups passed the test items than one would have expected from the standardization norms. The only possible explanation for this finding seems to be that only forty items were given in the testing period of fifty minutes. This allowed the children to pass many items they would not have been able to pass under speeded test conditions. It was also found in this experiment that when the test items were read orally to lower-class children while they followed in their test booklets, they passed appreciably more of the items ($P = .07$) than matched groups of children who took the test in the traditional (silent) manner (see Appendix). Such findings suggest that a wide range of conditions exist in our present tests and testing procedures which serve to introduce bias into our present measures of intelligence.

Finally, the emphasis on speed (rather than power) in the measurement of intelligence actually results in the confounding of such factors as reading speed, previous familiarity with the content of the test items, and rote or incidental memory with problem-solving ability. In attempting to develop unbiased tests, it seems desirable to remove such sources of bias from intelligence test scores, especially since previous experience with test-type materials is enjoyed differentially by various groups in our society, and the correlation between incidental memory and problem-solving ability is negligible, if indeed these two variables are not negatively related (cf. 31).

SUMMARY

"The standard-type intelligence tests are inadequate on several counts. Among other things, (a) they have measured only a very narrow range of mental abilities, namely those related to verbal or academic success, and have ignored many other abilities and problem-solving skills, which are perhaps more important for adjustment and success—even in middle-class society; (b) they have failed to provide measures of the wide variety of qualitative differences in the modes or processes of solving mental problems; (c) they have ignored the influences of differences in cultural training and socialization on the repertoire of experience and the attitude, motivation, and personality patterns of sub-groups in our society, and the effect of such factors on mental test performance; and (d) they have considered mental functioning in isolation, thus ignoring the interdependence of the individual's motivational and personality structure on the characteristics of his mental functioning, as seen, for example, in the differences between rote learning and the ability to use previous experiences creatively in new contexts.

"A re-evaluation of the purposes and problems involved in the appraisal and description of mental abilities is necessary before adequate mental tests can be developed. But before this can be done, it will first be necessary to conduct anthropological, sociological, and psychological studies to learn how representative children in our society live. For lower-class and ethnic children, for example, information is needed concerning their value, attitude, and motivational systems, the nature of their daily experiences, and the range of mental behaviors and modes of thinking used in finding solutions to their life problems. It will also be necessary to consider the growing body of evidence that mental functioning does not exist in a vacuum, but that the individual's motivational and personality structure, his attitudes, interests, needs, and goals are intimately related to, and in a large measure determine, his mental processes" (17).

REFERENCES

1. Ataullah, Kaniz. Cultural influence on children's solution of verbal problems: A qualitative study. Unpublished Ph.D. dissertation, University of Chicago, 1950.
2. Bentley, A. F. Kennetic inquiry. *Science*, 1950, 112, 775–783.
3. von Bertalanffy, L. The theory of open systems in physics and biology. *Science*, 1950, 111, 23–29.
4. Binet, A. & Simon, Th. *The development of intelligence in children* (Tr. by Elizabeth S. Kite). Baltimore: Williams & Wilkins, 1916.
5. Boas, F. Evidence on the nature of intelligence furnished by anthropology and ethnology. *Addresses and discussions presenting The Thirty-Ninth Yearbook, Intelligence: Its nature and Nurture, NSSE*, (Ed. by G. M. Whipple) Salem, Mass.: Newcomb and Gauss, 1940.
6. Bohr, N. On the notions of causality and complementarity. *Science*, 1950, 111, 51–54.
7. Cantril, H., Ames, A., Jr., Hastorf, A. H., & Ittelson, W. H. Psychology and scientific research: I. The nature of scientific inquiry; II. Scientific inquiry and scientific method; III. The transactional view in psychological research. *Science*, 1949, 110, 461–464, 491–497, 517–522.
8. Davis, A. *Social class influences upon learning* (The Inglis Lecture, 1948). Cambridge, Mass.: Harvard Univ. Press, 1948.
9. Davis, W. A. & Havighurst, R. J. The measurement of mental systems. *Sci. Monthly*, 1948, 66, 301–316.
10. Davis, A. & Eells, K. *Manual for the Davis-Eells test of general intelligence.* Yonkers-on-Hudson, N. Y.: World Book Co. (In preparation.)
11. Davis, A., Eells, K., & Bergman, D. Reasoning processes underlying pupil's choices on a standard test of intelligence. Unpubl. typescript, Dept. of Education, Univ. of Chicago, 1950.
12. Dobzhansky, Th. Heredity, environment, and evolution. *Science*, 1950, 111, 161–166.
13. Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. *Intelligence and cultural differences.* Chicago: Univ. of Chicago Press, 1951.
14. Einstein, A. & Infeld, L. *The evolution of physics.* New York: Simon & Schuster, 1942.
15. Gordon, L. V. & Durea, M. A. The effect of discouragement on the revised Stanford-Binet scale. *J. genet. Psychol.*, 1948, 73, 201–207.
16. Gulliksen, H. *Theory of mental tests.* New York: Wiley & Sons, 1950.
17. Haggard, Ernest A. Social-status and intelligence: An experimental study of certain cultural determinants of measured intelligence. *Genet. psychol. Monogr.* (In Press, May, 1954.)
18. Harlow, H. F. The formation of learning sets. *Psychol. Rev.*, 1949, 56, 51–65.
19. Hebb, D. O. *The organization of behavior.* New York: Wiley & Sons, 1949.
20. Jones, L. V. A factor analysis of the Stanford-Binet at four age levels. *Psychometrika*, 1949, 14, 299–331.
21. Lantz, B. Some dynamic aspects of success and failure. *Psychol. Monogr.*, 1945, 59, 6–21.
22. Lorge, I. Schooling makes a difference. *Teach. Coll. Rec.*, 1945, 46, 483–492.
23. Maslow, A. H. Problem-centering vs. means-centering in science. *Philos. Sci.*, 13, 326–331.
24. McGee, J. W. Review of Eells, et. al. (See reference 13.) *Am. Cath. soc. Rev.*, 1952, 13, 45.
25. McNemar, Q. *The revision of the Stanford-Binet scale.* Boston: Houghton Mifflin Co., 1942.
26. McNemar, Q. *Psychological statistics.* New York: Wiley & Sons, 1949.
27. McNemar, Q. Review of Eells, et. al. (See reference 13.) *Psychol. Bull.*, 1952, 49, 370–371.

28. **Rothstein**, J. Information, measurement, and quantum mechanics. Science, 1951, 114, 171–175.
29. **Sacks, Elinor**, L. Intelligence scores as a function of experimentally established social relationships between child and examiner. *J. abn., soc. Psychol.*, 1952, 47 (No. 2, Suppl.), 354–358.
30. **Sargent**, S. S. Review of Eells, et. al. (See reference 13.) *Amer. J. Soc.*, 1952, 58, 209–210.
31. **Saugstad**, P. Incidental memory and problem-solving. *Psychol. Rev.*, 1952, 59, 221–226.
32. **Stone**, D. R. Certain verbal factors in the intelligence-test performance of high and low social status groups. Unpubl. Ph.D. dissertation, Univ. of Chicago, 1946.
33. **Terman**, L. M. *The measurement of intelligence.* Boston: Houghton Mifflin Co., 1916.
34. **Terman**, L. M., et. al. *The Stanford revision and extension of the Binet-Simon scale for measuring intelligence.* Baltimore: Warwick & York, 1917.
35. **Terman**, L. M., **Dickson**, V. E., **Sutherland**, A. H., **Franzen**, R. H., **Tupper**, C. R., & **Fernald, Grace**. *Intelligence tests and school reorganization.* Yonkers-on-Hudson, N. Y.: World Book Co., 1923.
36. **Terman**, L. M., & **Merrill, Maud** A. *Measuring intelligence.* Boston: Houghton Mifflin Co., 1937.
37. **Thorndike**, E. L., **Bregman**, E. O., **Cobb**, M. V., & **Woodyard, Ella**. *The measurement of intelligence.* New York: Bureau of Publications, Teachers College, Columbia University, 1927.
38. **Warner**, W. L. (Personal communication.)
39. **Watson**, D. L. *Scientists are human.* London: Watts & Co., 1938.
40. **Weisskopf, Edith** A. Intellectual malfunctioning and personality. *J. abn., soc. Psychol.*, 1951, 46, 410–423.

APPENDIX

From: SOCIAL-STATUS AND INTELLIGENCE

An Experimental Study of Certain Cultural Determinants of Measured Intelligence

ERNEST A. HAGGARD

(Genetic Psychology Monographs, In Press)

May, 1954

I. *Major Variables and Experimental Conditions:*

Social-Status: On the basis of ISC scores,* 671 subjects were selected from approximately the top 14 and bottom 14 per cent of all eleven-year-old children in a Midwestern city of 115,000. They are designated "high-status" and "low-status" groups.

Practice: Fifty minute periods of practice for three consecutive days, spent solving test problems (items) similar to those used in the Initial Test and the Retest. All children receiving "practice" finished all items in the workbooks provided during each of the three practice sessions.

Motivation: Promise of a free theater pass, or its equivalent in money, if the child "did his best" during the Practice or Retest sessions. All subjects in these groups were given this "reward" at the end of the Practice and/or Retest sessions.

Form of Test Items: There were two parallel forms of 40 items each; the "Standard," taken from published intelligence tests; and the "Revised," which were rewritten as, for example:

"Standard" item	*"Revised" item*
Cub is to bear as gosling is to	Puppy goes with dog like kitten goes with
1 ( ) fox, 2 ( ) grouse, 3 ( ) goose, 4 ( ) rabbit, 5 ( ) duck.	1 ( ) fox, 2 ( ) goose, 3 ( ) cat, 4 ( ) rabbit, 5 ( ) duck.

Selection of test items: Fourteen months prior to the present experiment, 2,295 nine- and ten-year old children were given a battery of three intelligence tests, and 2,510 thirteen- and fourteen-year-old children were given a battery of four intelligence tests. The ages of the children in the present experiment averaged 11 years 2.57 months and 11 years 2.76 months for the high- and low-status groups respectively.

Four criteria determined the selection of the 40 items used in this experiment, namely that (a) the content of the items selected was meaningful to the children of both social classes, and (b) that these items could be revised without changing the basic meaning or the difficulty of the mental task involved. In addition, (c) items were selected which, on the basis of the previous testing, were passed more often (i.e., $P < .01$) by the high-status than by the low-status children of the same age. In order to find items of suitable difficulty for the 11-year-olds in this study, items were selected which were (d) failed by most of the younger children, and passed by most of the older children in the previous testing. Approximately two-thirds of the 40 items were taken from tests given to the older age group in the previous testing.

Presentation of Items on the Revised Retest: Most of the Revised Retests were taken in the traditional manner, but two groups had the Revised Retest read orally by the teacher while the children followed in their test booklets.

---

* An index giving equal weight to: parental education, parental income, house type, and dwelling area.

II. *The Experimental Design* given below represents the breakdown for one social-status group. The design is identical for both social-status groups.

DAY 1	DAYS 2–4	DAY 5		NUMBER IN STATUS GROUPS	
INITIAL TEST	PRACTICE PERIODS (50 min. per day)	RETEST		HIGH	LOW
Standard	Practice with Motivation	Retest with Motivation	Revised	28	28
			Revised (Oral)	32	28
			Standard	24	21
		Retest with No Motivation	Revised	21	17
			Standard	19	20
	Practice with No Motivation	Retest with Motivation	Revised	24	19
			Standard	18	19
		Retest with No Motivation	Revised	25	21
			Standard	19	22
	No Practice	Retest with Motivation	Revised	26	26
			Standard	23	20
		Retest with No Motivation	Revised	20	26
			Standard	25	26
Revised Only				35	39
			Totals	339	332

III. *Matching of Subjects:* Within each social-status group, subjects were matched on: (a) ISC, (b) age to the nearest month, (c) grade in school, and (d) Kuhlmann-Anderson I.Q. The means for each of the 14 high-status and 14 low-status sub-groups deviated not more than one standard error from the mean of their respective total social-status groups on any one of these four variables.

IV. *Control Variables:* The following data were collected for each child and utilized in the general statistical analysis: the child's (a) ISC, (b) age, (c) grade, (d) I.Q., (e) sex, (f) school, (g) teacher of the practice periods, (h) score on the Initial Test; the presence or absence of (i) Practice, (j) Motivated Practice, (k) Motivated Retest; whether the retest was (1) Standard or Revised form, and if Revised, (m) whether it was administered silently or orally, (n) the score on the Retest; and (o) the gain in performance as indicated by the difference between the transformed Initial Test and the transformed Retest scores.

V. *Scores used in General Statistical Analysis:* The difference between the transformed (arc sine) Initial Test and Retest scores.

VI. *Methods of Data Analysis:* For the general statistical analysis the Johnson-Neyman method of testing linear hypotheses was used to test the possible effects of variables or experimental conditions. For item analyses, Chi-square (with Yates' correction for continuity) was used. For appraising the degree of relationship between variables, the product-moment correlation was used.

Techniques for the Development of Unbiased Tests

QUINN McNEMAR

DISCUSSION OF PAPERS

> **Editor's Note**: As Dr. McNemar had not had an opportunity to read Dr. Haggard's speech prior to the Invitational Conference, the following material was prepared after the conference. In addition it was agreed that Dr. Haggard would be allowed to prepare a reply to Dr. McNemar for inclusion in the *Proceedings.*

Professor Lorge has successfully anticipated Dr. Haggard's *general* thesis and provided us with such an excellent critical evaluation thereof that little is left for me to say except that I am in full agreement with all the points made by Lorge.

Dr. Haggard's rather over-lengthy presentation contains many matters of a *specific* nature which I would like to question but time permits me to consider only a few points. Indeed, an adequate assessment of parts of his paper must await the publication of a number of researches which he cites.

First, I would like to set the record straight regarding my supposed failure to differentiate between the concepts of reliability and validity. Dr. Haggard gives two quotations, from two of my publications, which seem so inconsistent as to make "it not clear whether a measure of validity, or a measure of reliability, was used in standardizing" the 1937 Stanford-Binet. Perhaps my supposed inconsistency can be removed by merely pointing out that the first quotation happens to be from an introductory chapter by Terman—certainly Terman did not need to agree with something I was to write years later!

I find it difficult to share Haggard's alarm about our psychological measurement being modeled after classical physics. In fact, I would be quite happy if our schemes for measuring behavior could reach the level of commonplace measurement attained by the classical physicists.

One may question the clarity of parts of Haggard's discussion. For example, what does it mean to say that item difficulty, reliability, and validity "have formed the essential justification for many tests of intelligence?" And where did he get the idea that these three concepts "are supposed to be kept separate?" He didn't find this strawman in the sound treatise by Gulliksen (2). Nor will our speaker find in Gulliksen or any other modern source the notion that test-retest with an interval of sixteen months is an acceptable way for determining reliability.

Haggard says that the "concept of validity has sometimes been treated even more casually than reliability," and as a first bit of evidence he gives a quotation from page 7 of Terman and Merrill (3) which presumably tells us how items were selected for the 1937 Stanford-Binet. Actually, the complete sentence from which the quotation was lifted speaks of how "types of test items" were selected—quite a different thing. Next he gives further quotations regarding item selection for the 1937 Stanford-Binet, but never a hint that these quotations are from a section dealing with the preliminary selection of items. Since Terman readily admits that the 1937 scale measures essentially what was measured by the 1916 scale, Haggard asks for evidence regarding the validity of the latter—his own search having conveniently ignored the literature between 1917 and 1937.

As to his discussion of the question of I.Q. constancy, I can only remark that Haggard attributes a far greater degree of constancy than test-retest facts warrant. It is of course convenient for his thesis to have constancy for the I.Q. otherwise he would have to explain how continuation in the same social status level and continuing in "middle-class" schools could lead to changes in the I.Q.

The discussion of the reviews of Eells' book I find very amusing, especially since it purports to show how psychologists are reluctant to consider approaches which deviate from the orthodox while sociologists are more willing to accept the new. This deduction is arrived at by a strange type of logic: Eells' book received more favorable reviews in the sociological than in the psychological journals, ergo, Q.E.D. But this absurdity becomes ludicrous when it is noted that one of the two cited "sociological" reviews was by a psychologist (S. S. Sargent)! Further "reluctance" on the part of psychologists to accept the new (and thereby get into "the main stream of scientific thought") is cited by Haggard: an editor of a well-known psychological publication would not accept a research report from the Chicago group. Now I don't know the merits of this case but perhaps the editor was aware of the Bernardine Schmidt fiasco.

After an hour (or some 28 typescript pages) our speaker finally came to the question posed for discussion by this panel. No doubt some of you listened in vain for the detailed steps by which he proposes to develop unbiased tests. I found myself woefully confused at this juncture—earlier in his paper he criticised Terman for standardizing the Stanford-Binet on children from schools of average social status because that means "middle-class" schools, but now we learn that our "test constructor is a middle-class individual, being a professional person with college training." This equating of average social status with the college educated errs as much in the direction of imprecision as the concept of "six clearly marked social classes" (Eells et al., 1, p. 17) errs in the direction of pseudo precision.

Unfortunately Haggard's discussion does not permit one to evaluate the Chicago methods for developing unbiased tests—it is to be hoped that the cited forthcoming publications will provide the necessary detail for an appraisal. It will be interesting to learn the extent to which, and how well, these investigators are doing something different. Some of us will wish to know whether any of their methods lead to the elimination of the portion of the variance in test scores due to possible hereditary differences.

I have a few remarks to make on Dr. Rulon's paper. This test which he has devised is indeed very ingenious. I am sure that the feebleminded youngster will be glad to be tested by somebody who can, when giving directions, reach down to his level without the use of even the simplest words, thereby avoiding completely the vocabulary worries of our Chicago friends.

Rulon speaks of face validity, and to this I have no particular objection provided the notion isn't carried to the point where we delude ourselves. As I analyze this test, it seems to me that it involves a learning situation but at a higher conceptual level than the usual substitution type of stunt. Since this test is obviously a learning task, one must raise the question as to how general is the learning ability being tapped —the factor analysts may need to step in with an answer to this.

There is still another difficulty which Rulon must face. The learning theorist can ask whether performances on this learning task might be subject to transfer effects which are differential from person to person. Then the cultural protagonists can say that individuals in different cultures or in different social status levels will have learned different things, hence by way of possible transfer the influence of cultural differences may contribute to score variance.

Rulon claims that objects, actions, etc., required for this test are almost universal. Now I don't know whether the kids on the lower east side of New York City are familiar with cows; I rather doubt it. The hand-out illustration did not include the cow, but another illustration which I saw did. If the cow is used, I hope there is no sitting cow involved!

Although Rulon is properly cautious, he says that he thinks this test does not involve either visual acuity or perceptual ability. Those of us who have examined the illustrative material may think otherwise. I suspect that differences in perception may enter into performance on this test—as a positive suggestion for further eliminating visual and perceptual factors I suggest that in revising this scale he have the woman drawn by Peter Arno!

REFERENCES

1. EELLS, K., *et. al. Intelligence and cultural differences.* Chicago: Univ. of Chicago Press, 1951.
2. GULLIKSEN, H. *Theory of mental tests.* New York: John Wiley, 1950.
3. TERMAN, L. M., AND MERRILL, M. A. *Measuring intelligence.* Boston: Houghton Mifflin, 1937.

Techniques for the Development of Unbiased Tests

ERNEST A. HAGGARD

REPLY TO DR. McNEMAR'S REMARKS

Prefatory Note: It was understood when I accepted the invitation to appear on this Panel that the discussion was to be on a general theoretical level. In preparing my paper, I became more interested in the problem of bias in measuring intelligence than in confining my remarks to a time limit. Consequently, at the Conference, only part of the material was presented. Also, since I was asked to join the Panel relatively late, it was not possible to give Professor McNemar a copy of the paper before the Conference. Thus, it was later agreed that he be given an opportunity after the Conference to criticize my remarks. Professor McNemar's criticisms are based on his study of my paper over a six-week period. In replying to his criticisms, I will refer directly to them, paragraph by paragraph.

1. Regarding Professor Lorge's paper, I think he and I have approached the problem of "bias" in somewhat different manners, as is apparent from our papers. But in reacting to his comments, I would like to point out that a careful reading of Eells, et al. (2) will show that it is primarily a report of research investigating some of our present tests. The development of new intelligence tests is reported elsewhere (1). Also, in terms of Lorge's closing remarks, the ultimate purpose of the work by Davis and others at Chicago was to develop tests which "allow all in our democracy to have an equal opportunity for maximum development of their potentialities" because "some kinds of bias" have been removed from intelligence tests.

2. I had hoped that the time (six weeks) would permit McNemar to consider also some of the general methodological points I raised, especially since they are fundamentally more important in dealing with the problem of bias than the points he chose to discuss.

3. At the end of this paragraph, McNemar is correct in checking me up on the fact that Terman wrote Chapter I of his book (3). But my point was that this test was "validated" in a circular manner—and, indeed, it was rather a small circle—and that consequently this seems

to me more like a measure of reliability than of validity. (See paragraph 6 below.)

4. I am not alarmed, but rather believe that in the "measurement" of mental processes, the phenomena to be measured and the available means of "measuring" them differ from those of classical physics. Along with McNemar, I too "would be quite happy if our scheme for measuring behavior could reach the level of commonplace measurement attained by the classical physicists." If, however, "intelligence" somehow could be directly observed, and if we had scales which possess certain characteristics (e. g., equal units) *independent of* the phenomena being measured, we too could begin to make measurement statements in the manner of the classical physicist. But this is not the case, nor will wishing make it so.

5. I was only confessing my inability to see any other scientifically justifiable *raison d'être* (except for, say, prestige or monetary reasons) for some intelligence tests. No, I did not find in Gullikson the straw-man-idea that these concepts "are supposed to be kept separate." But this supposition is fairly common to our thinking—and McNemar's too—as seen in his desire "to set the record straight regarding my supposed failure to differentiate between the concepts of reliability and validity" (paragraph 3 above). Also, in many texts in this area, one finds such statements as "the familiar distinction between the 'reliability' of a test and its 'validity'" (5, 106). But my real point had to do with the inappropriateness of our conceptualization of our measurement problems, and hence the inappropriateness of various concepts or techniques that go along with, or fit, the conceptual model we use. Finally, I imagine that the reason test-retest reliabilities, with an interval of 16 months, are not "acceptable" is probably because they are generally too low to be of practicable value.*

6. There are several indications of this. In McNemar's book on the revision (3), he gives one chapter (VI) to a discussion of reliability, and, by his Index, pages 82-3 to validity--where, by the way, he says essentially what Terman said in Chapter I of his book. Furthermore, I do not believe that the selections I cited from *Measuring Intelligence* (4, 7-10) do violence in describing Terman's procedure. Item types which are eliminated in the preliminary screening certainly do not

---

* Actually, I see no theoretical reason why such a measure would not be acceptable, since mental age is presumed to grow at a rather steady rate, and since the correlation coefficient does not reflect differences between distribution means but only relative position within the two distributions.

appear in the final test, and from what I can gather, the procedure quoted from pages 9-10 is the same as the one used in deriving the final scales (4, 21-23).*

Incidentally, the purpose of the Panel was not to analyze the Stanford-Binet or review the literature, except as it pertains to the development of unbiased tests. But I want to take this opportunity to say that the 1937 Revision, if used wisely and skillfully is, pragmatically speaking, a very flexible and useful measuring and diagnostic instrument. I was certainly not advocating that it be discarded; I was talking about the relation of standardization procedures to possible sources of bias in measuring such phenomena as we call "intelligence."

7. I was quoting the work of Terman and others (cf. 5, 165), and did not argue for constancy of the I.Q. I did say, however, that more than social status level influences performance on intelligence tests.

8. I think that the tenor of McNemar's criticisms belies his amusement.

9. Now, McNemar knows that I confined my presentation to my allotted time, 30 minutes, and that the purpose of the Panel was a theoretical discussion of the problem of bias in tests. And as for my remark "by and large, we can be certain that the test constructor is a middle-class individual, being a professional person with college training," I was referring to a report of three studies which found that between 97.5 and 100 per cent of the public school teachers studied hold the values of middle-class or higher social status groups, (6, Ch. VIII). On the basis of such findings, I did not think that my generalization was unfair to the test constructors.

10. In reviewing McNemar's remarks about my paper, I am disappointed that his criticisms were not on a higher level, and that he failed to deal with some of the more fundamental issues raised. In view of the amount of time he had to "work over" my paper, I had hoped he would do more than concern himself with matters of wording and minor disagreements over incidental details. I trust that when McNemar "evaluates" and "appraises" forthcoming work in this area, he will use his abilities for the clarification of basic issues, and will do so in a manner in keeping with his stature in the field.

---

* For example, Terman says that "it was then possible to plot for each test the curve showing per cent of subjects passing in successive ages throughout the range. . . . The correlation of each test with composite total (equivalent to correlation with mental age) was computed separately for each test, thus providing a basis for the elimination of the least valid tests" (4, 22).

REFERENCES

1. See reference No. 10 above.
2. See reference No. 13 above.
3. See reference No. 25 above.
5. Goodenough, Florence L. *Mental testing.* New York: Rinehart & Co., 1949.
6. Warner, W. L., Havighurst, R. J., Loeb, M. B. *Who shall be educated?* New York: Harper & Bros., 1944.

DISCUSSION

PARTICIPANTS

RICHARD H. GAYLORD, ERNEST A. HAGGARD, PHILLIP J. RULON,
JOHN W. TUKEY

DR. GAYLORD: It seems to me we have two points of view in building these tests. One is that you sit down and build a test. It is going to be reasonably homogeneous in content. You then find out all the things that that kind of content is related to. I think, on the other hand, we have had the point of view that you take a lot of reference variables and find a test that is related to them in a predefined fashion. Those two positions are not compatible, you can't mix the two and come out with the same thing. I think each has its place.

QUESTION: I should like to hear Dr. Rulon defend himself on the question of the validity of his test.

DR. RULON: I don't remember having claimed any particular validity for the test.

I described this test to the Department of Psychology at Yale, and we had time for questions, so much time that I regret we do not have that situation today. But I thought the best question asked of me was the following: "Doctor, what are you going to say about the southern colored boy who doesn't do very well on your test?"

I said I would answer the question if the questioner would take my answer seriously. I didn't want to be accused of joking. The answer is, I shall say the child doesn't seem to be very good at this sort of thing.

DR. TUKEY: There are two or three questions I would like to raise. First, I take it that status and initial score have been confounded in this experiment, that is, the low status group on the whole has a lower initial score, and thus there is a question as to whether some of these differences may be due to the initial score position rather than status. (Beginning lower, they had a greater opportunity for increase!)

Second, there is a question I would like to ask for information. Essentially we have an analysis of variance here. Which error term was used for the conclusions?

Third, we seem to make statements separately about the low status and high status groups. Isn't the main interest of this operation the comparison of these groups, the interactions between status and other variables. If you look at the interactions, do they bear out all the conclusions that have been set down?

Dr. Haggard: As I understand confounding, it occurs when the effects of two or more variables, or treatments, or conditions, are thrown together, so that there is no means of identifying the source of variation attributable to each of them separately. This was not the case in this experiment. As I pointed out in the mimeographed Appendix, the subjects in both social-status groups were matched on four variables, and a number of other conditions were used as control variables. Each of these variables was used in the data analysis to partial out their separate effects in order to make more precise statements about the variable or condition under consideration. Consequently, even though in this experiment the low-status children averaged ten points lower on I.Q. and had an average of six months less schooling, this does not lead to confounding since the effects of each of the variables was controlled or accounted for.

As for your second question, I do not understand your statement that the error term was used for the conclusions. It is true that our significance tests are made up of a ratio, whether it is *t*, *F*, or Chi-square, in which the numerator is knowledge and the denominator is ignorance (or the error estimate)—that is to say, of the total variability among the data, the numerator is made up of the known or controlled sources of variation, and the denominator is the remainder, the unknown or uncontrolled sources of variation. One advantage of the Johnson-Newman technique is that the effects of such variables as school grade, I.Q., etc., are not left undetermined. Hence, the removal of the effects of the various known or controlled sources of variation from the denominator, or error estimate, serves to make the conclusions more precise.

Now, while I was making notes on your first two questions, you were asking a third, which I missed. Will you please ask it again?

Dr. Tukey: Apparently the conclusion is that a difference was significant by test for the high status children and the implication is that in the low status children it wasn't. What I would like to know, is the difference between high and low status children significant? Because it is quite possible to have, purely by chance, the difference for one status come out significant and the other not when the true difference is constant and the same.

Dr. Haggard: One point of my summary was that the condition of practice facilitated the gain in performance of the high-status children who took the standard form of the retest, and the gain of the low-status children who took the revised retest. On the standard retest, practice did help the high-status but not those from the low-status group.

Dr. Tukey: Did it have a negative effect, or non-significant effect?

Dr. Haggard: It was a non-significant effect.

Dr. Tukey: Have you any idea of the value?

Dr. Haggard: Not at the moment, except to say that the *P*-value fell below the .10 level. The *F*-value for the high-status group was 19.56 with 1 and 120 degrees of freedom.

Dr. Tukey: It is perfectly possible to get by reasonable sampling an F of 19.6 in one case and a non-significant F in another, where the population values are just the same, and so it seems to me the real question hasn't been answered in Point 1 at all. Are we sure there is a difference between the two groups in this characteristic?

Dr. Haggard: Although I did not mention the comparison you are asking for in my summary statement, it is in the monograph in press (17). It reads, "In fact, the high-status groups profited significantly more from the Practice Sessions than did the low-status children ($F1{,}240 = 7.32$; $P < .01$) when the Standard form of the Retest was given."

Appendix

Participants—1952 Invitational Conference on Testing Problems

Afflerbach, Janet, American Public Health Service
Ahmann, J. Stanley, Cornell University
Allen, Charles, Educational Testing Service
Allison, Roger, Educational Testing Service
Alman, John E., Boston University
Alt, Pauline M., Teachers College of Connecticut
Anastasi, Anne, Fordham University
Anderson, Roy N., North Carolina State College
Andrews, T. G., University of Maryland
Angoff, William, Educational Testing Service
Anthony, C. William, Maryland State Department of Education
Appel, Valentine, Richardson, Bellows, Henry & Co., Inc.
Armstrong, Fred, Lehigh University
Arnold, Samuel T., Brown University
Arsenian, Seth, Springfield College
Ash, Philip, Inland Steel Company
Avakian, Rose, Columbia University
Ayer, Frederic L., Teachers College, Columbia University
Baier, Donald E., Personnel Research Section, Adjutant General's Office
Bannon, Charles J., Crosby (Connecticut) High School
Barnes, Paul J., World Book Company
Bartnik, Robert V., Educational Testing Service
Baxter, Brent N., Prudential Insurance Company
Bayroff, A. G., Adjutant General's Office
Bear, Robert M., Dartmouth College
Beardsley, Katharine, Barnard College
Beck, Hubert P., City College of New York
Becker, Theodore, Civil Service, Albany
Beeler, Nelson F., Teachers College, Potsdam, New York
Belt, Sidney, Educational Testing Service
Bennee, M., Teachers College, Columbia University
Bennett, George K., The Psychological Corporation
Benson, Arthur, Educational Testing Service
Berdie, Ralph F., University of Minnesota
Berger, Bernard, Municipal Civil Service Commission, New York
Bergesen, B. E., Personnel Press, Inc.
Berne, Ellis J., State Department of Civil Service, Albany
Berne, Gerda, Albany, New York
Bernstein, Alvin J., Teachers College, Columbia University
Bigley, Maureen J., Archdiocesan Vocational Service
Bishop, Robert W., University of Cincinnati
Bishop, Ruth, National League for Nursing
Bittner, Reign H., Prudential Insurance Company
Blackman, Ruth, Educational Testing Service
Blaul, R. Elizabeth, Highland Park High School, Illinois
Boasi, Veronica M., Archdiocesan Vocational Service
Bollenbacher, Joan, Cincinnati Public Schools
Bowles, Frank H., College Entrance Examination Board
Braca, Susan E., Archdiocesan Vocational Service
Brandt, Hyman, American Occupational Therapy Association
Bray, Douglas W., Columbia University
Bretnall, William, Educational Testing Service
Bridges, Claude F., World Book Company
Brolyer, Cecil, Civil Service, Albany
Brooks, Douglas, Harvard University
Brooks, Richard B., College of William and Mary

TESTING PROBLEMS

BRYAN, Miriam M., Silver Burdett Company
BRYAN, Ned, Rutgers University
BUCK, Julia, Educational Testing Service
BUCKINGHAM, Guy, Allegheny College
BUCKTON, LaVerne, Brooklyn College
BURKE, James M., Darien Public Schools
BURKE, Paul J., Columbia University
BURNHAM, Paul S., Yale University
BUROS, Oscar K., Rutgers University
BYRNE, Richard H., University of Maryland
CANER, Faruk, The Turkish Embassy
CAPPS, Marian P., Howard University
CARLSON, C. Ray, Maxwell Air Force Base, Alabama
CARLSON, Harold, Upsala College
CARROLL, John B., Harvard University
CARSTATER, Eugene, Bureau of Naval Personnel
CARY, James L., Howard University
CASE, Ethel E., Polytechnic Institute of Brooklyn
CAYNE, Bernard, Educational Testing Service
CHAUNCEY, Henry, Educational Testing Service
CHURCHILL, Ruth D., Antioch College
CLARK, Priscilla, Educational Testing Service
CLENDENEN, Dorothy, University of California at Los Angeles
COBB, William E., Pennsylvania State College
COFFMAN, William, Educational Testing Service
COHEN, Joseph, City College of New York
COLADARCI, Arthur, Stanford University
COOMBS, Clyde H., University of Michigan
COPELAND, Herman A., Atlantic Refining Company
CORCORAN, Mary, Educational Testing Service
CORNELL, Ethel L., State Department of Education, Albany
COWLES, John T., Educational Testing Service
COX, H. M., University of Nebraska
COY, Genevieve, The Dalton School
CRANE, Harold, Educational Testing Service
CRANE, Percy F., University of Maine
CRAWFORD, J. R., University of Maine
CRISSY, William J. E., Queens College
CROOK, Frances, World Book Company
CURRAN, Florence, Educational Testing Service
CYNAMON, Manuel, Brooklyn College
DAHNKE, Harold, Michigan State College
DALY, Alice T., New York State Education Department
DAVIDSON, Helen H., City College of New York
DAVIS, Frederick B., Hunter College
DAVISON, Hugh M., Pennsylvania State College
DERRICK, Clarence, Educational Testing Service
DETCHEN, Lily, Pennsylvania College for Women
DIAMOND, Lorraine K., City College of New York
DIEDERICH, Paul, Educational Testing Service
DIERS, Helen A., Vocational Advisory Service
DION, Robert, California Test Bureau
DIVESTA, Francis J., Maxwell Air Force Base, Montgomery, Alabama
DOBBIN, John, Educational Testing Service
DOPPELT, Jerome E., The Psychological Corporation
DRAGOSITZ, Anna, Educational Testing Service
DRAKE, Lewis E., University of Wisconsin
DRESSEL, Paul L., Michigan State College
DRY, Raymond J., Life Insurance Agency Management Association
DUNN, Catherine, Educational Testing Service
DUNSTAN, William H., Schenectady Public Schools
DUROST, Walter N., Test Service & Advisement Center
DYER, Henry S., College Entrance Examination Board
EBEL, Robert L., State University of Iowa
EDWARDS, Robert, New York State Psychiatric Institute
EL-KOUSSY, Abdul Aziz, Institute of Education, Cairo, Egypt
ENGELHART, Max D., Chicago City Junior Colleges
EPSTEIN, Bertram, City College of New York
FAN, C. T., Educational Testing Service
FATERSON, Hanna F., State University of New York
FAY, Paul J., State Department of Civil Service, Albany, New York
FERGUSON, Leonard W., Aetna Life Insurance Company

Findley, Warren, Educational Testing Service
Finkle, Robert B., Metropolitan Life Insurance Company
Flanagan, John C., American Institute for Research
Fleming, Mae, Educational Testing Service
Flemming, Edwin G., Burton, Bigelow Organization
Foss, Ciara R., American Psychological Association
Fouratt, Jean, Educational Testing Service
Fox, William H., Indiana University
Fredriksen, Norman, Educational Testing Service
Freeman, Paul, Educational Testing Service
Freidman, Sidney, Bureau of Naval Personnel
French, Benjamin J., State Department of Civil Service, Albany, New York
French, John, Educational Testing Service
Frutchey, Fred P., U. S. Department of Agriculture
Fuchs, Edmund F., Personnel Research Section, Adjutant General's Office
Furst, Edward J., University of Michigan
Gallagher, Henrietta, Educational Testing Service
Gardner, Eric F., Syracuse University
Garrett, Henry E., Columbia University
Gaylord, Richard H., Personnel Research Section, Adjutant General's Office
Gehlmann, Frederick, Science Research Associates
Gekoski, Norman, Temple University
Gelink, Marjorie, The Psychological Corporation
Gerberich, J. Raymond, University of Connecticut
Giangrande, Salvatore C., St. John's University
Goodman, Samuel M., HRRI Maxwell Field, Alabama
Green, Bert F., Massachusetts Institute of Technology
Greene, Harry A., Iowa State
Grimm, Elaine R., American Public Health Association
Gulliksen, Harold, Educational Testing Service
Gustad, John W., University of Maryland
Haddad, R. K., Brooklyn College
Hagen, Elizabeth P., Teachers College, Columbia University
Haggard, Ernest A., University of Chicago
Haggerty, Helen R., Personnel Research Section, Adjutant General's Office
Hagman, Elmer R., Greenwich Public Schools
Hana, Attia M., Teachers College, Columbia University
Harmon, Kathryn, Educational Testing Service
Harmon, Leon, Princeton, New Jersey
Harris, David, Educational Testing Service
Hart, May E., Babcock & Wilcox
Hastings, J. Thomas, University of Illinois
Hedlund, Paul A., State Department of Education, Albany, New York
Helmick, John, Educational Testing Service
Hieronymus, Albert N., State University of Iowa
Hoberman, Solomon, Municipal Civil Service Commission
Hoffman, E. Lee, Tulane University
Hogarth, Rhoda, Parent-Child Consultation Center
Horowitz, Milton W., Queens College
Horton, Clark W., Dartmouth College
Howard, Robert West, Annisquam, Massachusetts
Huddleston, Edith, Educational Testing Service
Hunt, Barbara, Educational Testing Service
Hyman, Herbert, Columbia University
Jacobs, Robert, A & M College of Texas
Jaspen, Nathan, National League for Nursing Education
Jeffrey, Wendell E., Barnard College
Jenkins, Thomas N., New York University
Jennings, Helen H., Brooklyn College
Johnson, A. P., Educational Testing Service
Jones, Edward S., University of Buffalo
Jones, Vernon, Clark University
Kaback, Goldie Ruth, City College of New York
Karon, Bertram P., Educational Testing Service
Kernan, John P., Dunlap Associates
Kelley, DeCourcy, Educational Testing Service
Kerr, Colin H., Boston University Junior College
Kidd, John W., Michigan State College

Kimball, Elizabeth, Educational Testing Service
King, Richard G., Harvard University
Kipnis, David, Richardson, Bellows, Henry & Co.
Kirkpatrick, Forrest H., Bethany College
Kline, William E., The Choate School
Kogan, Leonard S., Community Service Society
Kolkebeck, Robert F., Educational Testing Service
Kushner, Rose E., City College of New York
Lamke, T. A., Teachers College, Iowa
Langmuir, Charles R., Syracuse University
Lannholm, Gerald, Educational Testing Service
Layton, Wilbur L., University of Minnesota
Lazo, Elizabeth, American Public Health Service
Leach, Kent W., University of Michigan
Leach, Sarah C., The Psychological Corporation
Lembke, Glenn L., Randolph Air Force Base, Texas
Lennon, Roger T., World Book Company
Lev, Joseph, State Department of Civil Service, Albany
Leverett, Hollis M., American Optical Company
Levine, Richard, Educational Testing Service
Levy, Charlotte National League for Nursing
Lincoln, A. L., Lawrenceville School
Lindquist, E. F., State University of Iowa
Long, Lillian D., Professional Examination Service
Long, Louis, City College of New York
Lord, Fred, Educational Testing Service
Lord, Shirley, Educational Testing Service
Lorge, Irving, Teachers College, Columbia University
Lorr, Maurice, Veterans Administration
Lucas, Charles, Educational Testing Service
Lurie, Walter A., National Community Relations Advisory Council
Lutz, Orpha L., Montclair State Teachers College
McArthur, Charles C., Harvard University
McCall, W. C., University of South Carolina
McCambridge, Barbara, Educational Testing Service
McCann, Forbes E., Civil Service, Albany
McClelland, David C., Wesleyan University
McCollum, Joyce E., Civil Service, Albany
McCully, C. Harold, Veterans Administration
McFarland, Henry, Princeton University
McGuire, John P., State Education Department, Albany
McKinney, Lidie, Educational Testing Service
McNemar, Quinn, Stanford University
McQuitty, John V., University of Florida
MacPhail, Andrew H., Brown University
Malcolm, Donald, Maxwell Air Force Base, Montgomery, Alabama
Manuel, Herschel T., University of Texas
Marquis, Lloyd D., Educator's Washington Dispatch
Marsh, Donald D., Pace College
Marston, Helen, Educational Testing Service
Martin, Lycia O., Trenton State Teachers College
Mason, Frederic, University of Malaya
Mathews, Chester, Ohio Wesleyan University
Melville, S. Donald, Educational Testing Service
Merry, Robert W., Harvard Business School
Messick, Betty, Pennsylvania Bell Telephone Company
Messick, Samuel, Educational Testing Service
Mettler, James, Educational Testing Service
Mikle, Stephen, Educational Testing Service
Miller, Peter, Educational Testing Service
Mitchell, Blythe, World Book Company
Mitzel, Harold E., City College of New York
Moffat, Stanley, Educational Testing Service
Mollenkopf, William, Educational Testing Service

MORGAN, Henry H., The Psychological Corporation
MORRISON, Alexander, Brooklyn Polytechnic Institute
MUIRHEAD, Peter P., State Department of Education, Albany
MYERS, Charles, Educational Testing Service
NEWMAN, Sidney H., U. S. Public Health Service
NOLAN, Edward G., Educational Testing Service
NORTH, Robert D., University of Kentucky
NUCKOLS, Robert C., Life Insurance Agency Management Association
NULTY, Francis, Educational Testing Service
NYSTROM, Christine, Educational Testing Service
O'CONNOR, Virgil J., Headquarters, United States Air Force
O'KANE, Marianne, Educational Testing Service
OLSEN, Marjorie, Educational Testing Service
O'NEIL, William M., University of Sydney
ORLEANS, Issak D., Mitchel Air Force Base
ORLEANS, Joseph B., George Washington High School
ORSHANSKY, Bernice, Mitchel Air Force Base
ORSI, Antoinette, Educational Testing Service
OXTOBY, Toby, Commission on Human Resources
PACE, C. Robert, Syracuse University
PASHALIAN, Siroon, New York University
PEARSON, Richard, Educational Testing Service
PERLMAN, Mildred, New York City Civil Service Commission
PERLOFF, Robert, Personnel Research Section, Adjutant General's Office
PERRY, William D., University of North Carolina
PETERSON, Donald A., Life Insurance Agency Management Association
PETERSON, Shailer, American Dental Association
PETERSON, William C., New Jersey Standard Oil Company
PHILLIPS, Laura M., Silver Burdett Company
PHILP, Hugh, Harvard University
PINZKA, Charles, Educational Testing Service
PLUMLEE, L. B., Educational Testing Service
POLIN, A. Terrance, Teachers College, Columbia University
POTTS, Edith M., The Psychological Corporation
PRESCOTT, George A., World Book Company
PRESTON, Braxton, Educational Testing Service
PRESTWOOD, Elwood L., Teachers College, Columbia University
QUICK, Robert, American Council on Education
RABINOWITZ, William, Division of Teacher Education, New York
RAPPARLIE, John H., Owens-Illinois Glass Company
RASKIN, Evelyn, Brooklyn College
REMMERS, H. H., Purdue University
REPPERT, Harold C., Temple University
REUTER, William, Educational Testing Service
RHULE, Warren, Educational Testing Service
RICCIUTI, Henry, Educational Testing Service
RICHARDSON, Marion W., Richardson, Bellows, Henry and Company
RICHARDSON, Ruth P., Richardson, Bellows, Henry and Company
RICHEY, Katherine, Rutgers University
RICKS, James H., The Psychological Corporation
RIMALOVER, Jack K., Educational Testing Service
RIMOLDI, Horacio, Educational Testing Service
RIVLIN, Harry N., Queens College
ROBBINS, Irving, Queens College
ROCA, Pablo, Department of Education, Puerto Rico
ROCK, Robert T., Fordham University
RULON, P. J., Harvard University
RUMMEL, J. Francis, University of Oregon
SADACCA, Robert, Educational Testing Service
SAIT, Edward, Rensselaer Polytechnic Institute
SANDERS, Edward, College Entrance Examination Board
SANFORD, R. Nevitt, Vassar College
SARGENT, S. Stansfeld, Barnard College
SAUNDERS, David, Educational Testing Service
SCATES, Alice Yeomans, American Council on Education

Scates, Douglas, American Council on Education
Schaefer, Willis C., Institute for Research in Human Relations
Schreider, Rose, Educational Testing Service
Schrader, W. B., Educational Testing Service
Schulman, Hugh, Queens College
Schultz, Douglas, Educational Testing Service
Schultz, Margaret, Educational Testing Service
Schweiker, Robert F., Educational Research Corporation
Scribner, Peter C., World Book Company
Seashore, Harold G., The Psychological Corporation
Segel, David, U. S. Office of Education
Selover, Robert B., Prudential Insurance Company
Sensenig, Herbert, Dartmouth College
Sharp, Catherine G., Educational Testing Service
Shaycoft, Marion F., American Institute for Research
Shienbloom, Charles, Board of Public Education, Philadelphia
Shields, William S., U. S. Naval Academy
Shutt, Charles N., Berea College
Smith, Alexander F., University of Connecticut
Smith, M. Brewster, Social Science Research Council
Smith, Muriel, Educational Testing Service
Solomon, Herbert, Teachers College, Columbia University
Solomon, Robert, Educational Testing Service
Souther, Mary T., Tower Hill School
Spaney, Emma, Queens College
Spaulding, Geraldine, Educational Records Bureau
Spencer, Douglas, U. S. Military Academy
Spencer, Lyle M., Science Research Associates
Stallman, F. B., New York Telephone Company
Stephan, Frederick F., Princeton University
Sternberg, Jack J., Queens College
Stewart, Naomi, Educational Testing Service
Stockhamer, Nathan N., Teachers College, Columbia University
Stokes, Thomas M., Metropolitan Life Insurance Company
Stone, Paul T., Huntington College
Stouffer, Samuel A., Harvard University
Stoughton, Robert W., State Department of Education, Connecticut
Sullens, Reginald H., American Dental Association
Swanson, Edward, Educational Testing Service
Swineford, Frances, Educational Testing Service
Symonds, Percival M., Teachers College, Columbia University
Symonds, Mrs. Percival M., New York City
Taber, Victor A., State Education Department, New York
Tasso, Charles A., Richardson, Bellows, Henry & Company
Tatsuoka, Maurice, Harvard University
Taylor, Calvin W., National Research Council
Taylor, Elsie, Educational Testing Service
Taylor, Justine, Educational Testing Service
Tchorni, Bernard, Educational Testing Service
Terral, J. E., Educational Testing Service
Thibault, Paula, Educational Testing Service
Thompson, Albert S., Teachers College, Columbia University
Thornton, Richard F., Fordham University
Tiedeman, David V., Harvard University
Tinkle, J. W., Mitchel Air Force Base, New York
Travers, Robert M. W., Human Resources Research Center
Traxler, Arthur E., Educational Records Bureau
Triggs, Frances, Commission on Diagnostic Reading Tests
Tucker, Ledyard R, Educational Testing Service
Tukey, John W., Princeton University
Tukey, Mrs. John W., Princeton, New Jersey
Turnbull, William, Educational Testing Service
Upshall, Charles C., Eastman Kodak Company
Wadell, Blandena C., World Book Company

Wagner, E. Paul, Teachers College, Bloomsburg, Pa.

Walker, Helen M., Teachers College, Columbia University

Wallace, Wimburn L., The Psychological Corporation

Wandt, Edwin, City College of New York

Wantman, Morey J., University of Rochester

Watkins, Richard, Pennsylvania State College

Weisluch, Edith G., Kings Point, New York

Weisz, Sylvia, Educational Testing Service

Weitz, Anne, Life Insurance Agency Management Association

Weitz, Joseph, Life Insurance Agency Management Association

Wellck, A. A., University of New Mexico

Wenzel, Bernice M., Barnard College

Wesman, Alexander G., The Psychological Corporation

Whitla, Dean, Harvard University

Whitmore, Howard S., New York Telephone Company

Whitney, Alfred G., Life Insurance Agency Management Association

Wicoff, Evelyn, Educational Testing Service

Wilcox, Glenn W., Boston University Junior College

Wilke, Marguerite M., Greenwich Public Schools, Connecticut

Wilke, Walter H., New York University

Wilks, S. S., Princeton University

Williams, Malcolm J., U. S. Coast Guard Academy

Wilson, Kenneth M., Harvard University

Winans, S. David, New Jersey Department of Education

Windle, Leon, Educational Testing Service

Wingo, Alfred L., Virginia State Department of Education

Winkler, Lila, Riverside Hospital

Wolfle, Dael L., Conference Board of the Associated Research Councils

Wolman, Benjamin, City College of New York

Wood, Ray G., Ohio State Department of Education

Wright, Wilbur H., Geneseo State Teachers College

Wrightstone, J. Wayne, Board of Education, New York

Zubin, Joseph, Columbia University